Bolt Beranek and Newman Inc.

BBN Report No. 3794

Speech Compression and Evaluation

R. Viswanathan, J. Makhoul, A.W.F. Huggins

April 1978

Prepared for:
Defense Advanced Research Project Agency

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: BBN [redacted] 3794	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): SPEECH COMPRESSION AND EVALUATION,	5. TYPE OF REPORT & PERIOD COVERED: Final Report. Dec [redacted]74 - Dec [redacted]77,
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): R. Viswanathan, J. Makhoul [redacted] A.W.F. Huggins	8. CONTRACT OR GRANT NUMBER(s): MDA 903-75-C-0180, ARPA Order-2935
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Bolt Beranek and Newman Inc. 50 Moulton St. Cambridge, MA 02138	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE: Apr[redacted]978
12 245p.	13. NUMBER OF PAGES: 228
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce for sale to the general public.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This research was sponsored by the Advanced Research Projects Agency under ARPA Order No. 2935.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech compression, Vocoders, Linear Predictive Vocoders, Linear Prediction, Covariance Lattice Methods, Adaptive Lattice Methods, Linear Predictive Warping, Log Area Ratios, Quantization, Spectral Sensitivity Analysis, Variable Rate Transmission, Perceptual Model of Speech, Optimal Linear

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

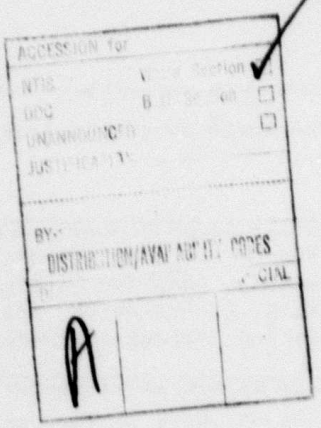
This report describes our work in the past three years on data compression and quality evaluation of digital speech. We developed and implemented linear predictive coding (LPC) techniques with the overall objective of digitally transmitting high quality speech at the lowest possible average data rates over packet-switched communication media. Major techniques reported include: covariance lattice method of linear prediction analysis, adaptive lattice methods, linear predictive spectral warping, improved

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

19. Key words (Cont.)
Interpolation, Mixed-Source Model, Robust Source Model, Speech quality evaluation, Subjective quality tests, multidimensional scaling, MDPREF, INDSCAL, Phoneme-specific intelligibility test, sequence effects in quality assessment, Objective Speech Quality Evaluation, Spectral Distance, Real-Time Implementation.

20. Abstract (Cont.)
quantization of LPC parameters, variable frame rate transmission of LPC parameters based on a functional perceptual model of speech, and a mixed-source model for LPC synthesizer to produce more natural-sounding speech. Also, we developed a reliable method for measuring subjective speech quality. This method was employed to formally demonstrate the quality improvements provided by our speech analysis/synthesis techniques as well as for studying speech quality as a function of LPC parameters. As subjective procedures are generally expensive and time-consuming, we developed and tested several objective procedures for speech quality evaluation. The results from these objective procedures were found to be highly correlated to the corresponding subjective quality judgments. Another highlight of our work is the development of a speech processing computer facility with the ultimate goal of transmitting narrowband speech in real time over the ARPA Network.



BBN Report No. 3794
April 1978

SPEECH COMPRESSION AND EVALUATION

Final Report
Contract No. MDA 903-75-C-0180
December 1974 - December 1977

R. Viswanathan
J. Makhoul
A.W.F. Huggins

Submitted to:

Defense Advanced Reserach Projects Agency

PROJECT PERSONNEL

John Makhoul	Principal Investigator
R. (Vishu) Viswanathan	Principal Investigator
A.W.F. (Bill) Huggins	Senior Scientist
Lynn Cosell	Research Engineer
William Russell	Research Engineer
Richard Schwartz	Research Engineer
Kathleen Starr	Project Secretary

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

1. INTRODUCTION

In this report, we present the results of our work in the past three years on data compression and quality evaluation of digital speech. The overall goal of our research has been to develop and implement techniques for digitally transmitting high quality speech at the lowest possible data rates. We have developed these techniques for Linear Predictive Coding (LPC) systems (also known as LPC vocoders). Also, they have been designed for transmitting speech over packet-switched communication media, an example of which is the ARPA Network; these media handle data messages in a time-asynchronous fashion. As a result, the data rate of our digital vocoder varies in time in accordance with the properties of the incoming speech signal. The variable transmission rate has a low upper bound as well as a low average, an important consideration for a real-time application such as transmission over the ARPA Network.

1.1 Summary of Major Results

Analysis Methods

We developed a new analysis method for linear prediction, called covariance lattice method. The method combines all the desirable properties of the traditional autocorrelation and covariance methods, and requires about the same computational complexity as the other two. These properties are: (1) Windowing

of the signal is not required; (2) The resulting all-pole linear prediction filter is guaranteed to be stable; (3) Stability is less sensitive to finite wordlength computations; and (4) Quantization of the lattice model parameters (for the purpose of data compression) can be accomplished within the recursion for retention of accuracy in representation.

We extended the lattice method to perform adaptive analysis in the sense of providing new estimates for the lattice parameters for every speech sample. Adaptive methods in general offer several advantages over the above "block" analysis methods; these include the option to choose which set of coefficient estimates to transmit in a given segment of the signal, and simpler hardware realization. In addition, our adaptive lattice methods ensure filter stability, and possess a desirable convergence property in that the convergence is almost independent of the spectral dynamic range of the input signal.

Also, we developed a linear predictive spectral warping technique to be included as part of the analyzer. This technique makes more effective use of the bits needed to transmit spectral information.

Parameter Quantization

We developed improved quantization schemes for LPC parameters: log area ratios (LARs), pitch and gain. The scheme for LARs employs unequal quantization step sizes for the different coefficients, with the step sizes derived by taking advantage of the differences in spectral sensitivity levels of individual LARs. The pitch quantization scheme makes efficient use of all the levels, in the sense that the decoded pitch values corresponding to these levels are all distinct. As LPC gain parameter, we found the energy of the speech signal to be a desirable choice.

Perceptual Model of Speech and Variable Frame Rate Transmission

We formulated and experimentally validated a functional perceptual model of speech in which speech is represented, with only a minimal loss in perceived quality, in terms of LPC parameters extracted time-asynchronously at a minimum set of time instances and in terms of linear parameter variation over the interval between these time instances. Based on this model, we developed new variable frame rate (VFR) transmission schemes for LARs, pitch and gain. We applied these VFR compression schemes to a 100 frames/sec fixed-rate LPC vocoder with a bit rate of about 5700 bps (bits/sec) to produce a variable rate vocoder with an average bit rate of only about 2100 bps for continuous speech

and with approximately the same speech quality as the fixed-rate system. Use of Huffman coding and variable order linear prediction (two of a number of techniques that we developed under a previous ARPA project [1]) would further lower the average bit rate to about 1500 bps, with no change in perceived speech quality.

A Mixed-Source Model to Improve Speech Synthesis

With the objective of enhancing the naturalness of the synthesized speech, we developed a new model for generating the excitation signal for the LPC synthesizer. In contrast to the traditional idealized pulse/noise (or voiced/unvoiced) source model, the new model mixes the pulse and noise excitations. The mix is achieved by dividing the speech spectrum into two regions, with the pulse source exciting the low-frequency region and the noise source exciting the high-frequency region. The cutoff frequency that separates the two regions is adaptively varied in accordance with the changing speech signal. Experiments using the new model indicated its power in synthesizing natural sounding voiced fricatives, and in largely eliminating the "buzzy" quality of vocoded speech.

Subjective Speech Quality Evaluation

We developed and tested an improved method for measuring subjective speech quality. The method uses a set of six specially designed sentences, each read by six talkers. The material is both representative, in that it covers a wide range of speech events and talker characteristics, and also challenging, in that some speech material is included that would fully extend any LPC vocoder's abilities. Applying this method, we obtained several practical results. For example, by studying speech quality as a function of vocoder parameters, we derived tradeoff relations to define the combination of vocoder parameters yielding the best quality for any desired overall bit rate. In another test, we showed that variable frame rate transmission techniques can produce the highest quality at any given rate, compared to two other methods which controlled the bit rate by adjusting the LPC order or by varying the log area ratio quantization step size. Also, we formally demonstrated the effectiveness of our perceptual-model-based VFR scheme and its superiority to our earlier log-likelihood ratio VFR scheme. In addition, we generated subjective speech quality data which we then used as a baseline for correlating against results obtained from our objective methods of speech quality assessment.

As part of our subjective speech quality work, we also investigated a few other topics including: (1) a phoneme-specific

intelligibility test, using nonsense materials; (2) the effect of lost packets on the intelligibility of speech transmitted over ARPANET; and (3) development of a method to reduce stimulus sequence effects on listeners' judgments.

Objective Speech Quality Evaluation

We formulated a general framework for objective speech quality evaluation of narrowband LPC vocoders. Within this framework, we developed several objective methods. In each method, the error in short-term spectral behavior between vocoded speech and the original is computed once every 10 ms. These errors are appropriately weighted and averaged over an utterance to produce a single objective score. We evaluated the objective methods by correlating the resulting objective scores with formal subjective speech quality judgments. The usefulness of our methods was clearly indicated by the high correlations that we obtained.

Real-Time Implementation

The current BBN speech facility has evolved mostly during the last three years. Briefly, it consists of the following: the SPS-41 computer with a dual-port memory interface and a dual channel A/D and D/A converter system; the PDP-11/40 computer with an RT11 operating system, an IMP11A interface to provide a link to the ARPA Network, the IMLAC PDS-1 display minicomputer as

a peripheral to the PDP11, and a software package which includes an FTP (File Transfer Protocol), a real-time speech acquisition, waveform display and editing program, and a convenient interactive playback program.

We cooperated with the other sites in the ARPA community in implementing an LPC vocoder that transmits speech over the ARPA Network in real time. Also, we provided specifications to ARPA LPC-II system, the first real-time variable-rate speech compression system on the ARPANET.

1.2 Outline of Report

Before we outline the contents of Sections 2-18, we note that the results of our work on various topics have been previously reported in the form of conference or journal papers and ARPA Network Speech Compression (NSC) notes. We describe these results briefly in the main sections of the report, and include these papers as appendices. Of course, topics that we have not previously reported, or on which additional work has been performed since the previous reporting, are dealt with in a detailed manner.

In Section 2, we describe three analysis methods: covariance lattice, adaptive lattice and linear predictive warping.

Section 3 contains the description of improved quantization schemes for LARs and pitch. Also considered in this section is the question of which of the two candidates for LPC gain parameter, speech signal energy and linear prediction error signal energy, produces a smaller quantization error.

Section 4 describes in detail our new variable frame rate transmission schemes for LARs, pitch and gain. First, we briefly review our work on VFR transmission performed on a previous ARPA project [1]. Then, we state our perceptual model of speech, and indicate a major difference between the previous VFR scheme and the new VFR scheme based on this perceptual model. Next, the various features of the new VFR scheme for LAR transmission are described at length, followed by the experimental results of comparisons of the speech quality of an LPC vocoder which transmitted LARs at a variable rate using this new scheme but pitch and gain at a fixed rate, with the speech quality of several other fixed-rate and variable-rate vocoders. Next, to substantially reduce the computational burden, we propose a simplified VFR scheme for LAR transmission. Finally, two types of VFR schemes for the transmission of pitch and gain are presented.

In Section 5, we consider our work on three issues related to the operation of the LPC synthesizer. These are: optimal linear interpolation of synthesizer parameters, implementation of synthesizer gain, and all-pass excitation.

Section 6 deals with our mixed-source model, an automatic scheme to extract the model parameter (cutoff frequency), implementation of the model at the synthesizer, and the effect on vocoded speech due to the use of the model.

Our work on subjective speech quality evaluation is presented in detail in Section 7. First, we describe the development and testing of a subjective quality measurement procedure. The results obtained by applying this procedure to three practical problems are given next. The section ends with discussions on several miscellaneous topics in the subjective quality evaluation area that we worked on as part of this project.

Section 8 deals with our efforts on the task of objective speech quality evaluation. The section starts with a statement of a general framework that we used in dealing with this task. Next, several distance measures are described for computing the error in short-term spectral behavior between vocoded speech and the original. Methods for time-weighting and time-averaging the computed frame spectral errors over an utterance are considered. Finally, the results obtained by comparing objective speech quality scores against subjective judgments are presented.

In Section 9, we describe our work towards developing a real-time speech facility at BBN. Also, we briefly summarize the

specifications that we provided for ARPA LPC-II speech compression system.

Two additional topics that we have also worked on during this project are considered in Section 9. These are: Differential Pulse Code Modulation (DPCM) coding of LPC parameters, and linear predictive formant vocoder.

2. ANALYSIS METHODS

A number of new analysis methods have been developed, some of which promise to have a major impact in various estimation and modelling applications. The first two sections describe our contributions to the area of lattice methods in linear prediction analysis. The last section presents the method of linear predictive spectral warping, which makes more effective use of the bits needed to transmit spectral information.

2.1 Covariance Lattice Methods

The autocorrelation method of linear prediction guarantees the stability of the all-pole filter, but has the disadvantage that windowing of the speech signal causes some unwanted distortion in the spectrum. In practice, even the stability is not always guaranteed with finite wordlength (FWL) computations. On the other hand, the covariance method does not guarantee the stability of the filter even with floating-point computation, but it has the advantage that there is no windowing and hence no unnecessary distortion of the signal spectrum. To combine the advantages of these two methods, we developed a new formulation for linear prediction, which we call the covariance lattice method (see Appendix 1 for details). The method is one of a class of lattice methods which guarantee the stability of the all-pole linear prediction filter, with or without windowing of

the signal, and with the number of computations being comparable to the autocorrelation and covariance methods; also, stability is less sensitive to FWL computations.

We incorporated the covariance lattice method into our floating-point simulation of the LPC speech compression system. This also involved "tuning" of such quantities as the analysis interval and the criterion for determining optimal LPC order. (The latter is required when variable order linear prediction is used [1].) The result was approximately the same speech quality as that from our earlier 1500 bps LPC system [1] (which used the autocorrelation method) at about the same total computation time. In fixed-point implementations, however, the lower sensitivity of filter stability to FWL computations provided by the covariance lattice method is expected to lead to an improvement in speech quality relative to that from the autocorrelation LPC system. Furthermore, the covariance lattice method permits the coefficients to be quantized within the recursion, thus integrating quantization into the coefficient estimation process; this is expected to improve the accuracy of the estimated short-term speech spectrum, and hence to improve the quality of the synthesized speech. (In non-lattice methods, quantization is done only after completing coefficient estimation.) However, one of the major benefits of lattice methods is expected to be in simpler hardware realizations.

2.2 Adaptive Lattice Methods

Covariance lattice methods are appropriate for "block" analysis of speech, whereby the speech is analyzed a frame at a time. However, for certain hardware realizations, it might be simpler to perform an adaptive type of analysis, which continuously updates the values of the reflection coefficients in the lattice. This has the advantage that one can choose which set of coefficients to transmit in a particular speech interval. Having such a choice might be important in obtaining consistent spectral estimates that are not as affected by the quasi-periodic nature of voiced speech as are the regular block estimation methods, such as the autocorrelation and covariance methods.

We have recently developed the theoretical basis for adaptive lattice estimation (see Appendices 2 and 3 for details). Although the methods have not been tested out thoroughly for speech, it is expected that they would give positive results. One of the major properties of adaptive lattice methods is that the convergence to the optimal values is almost independent of the spectral dynamic range of the input signal (i.e. independent of the eigenvalue spread of the signal covariance matrix). This property, absent in many previous adaptive methods, promises to have wide-ranging applications in communication systems, wherever adaptive transversal filters are used.

2.3 Linear Predictive Warping

In Appendix 4, we include a detailed description of a general method for the spectral distortion or warping of speech signals. The basic idea is to decompose the speech signal, on a short-time basis, into two components: a spectral envelope and an excitation signal. The spectrum is then warped in any desired manner and then recombined with the excitation to form a new signal with a warped spectrum but with the same pitch and intonation. The method has many potential applications, including unscrambling of helium speech, spectral warping for the hard-of-hearing, and more efficient communications.

The application to efficient communications is in the form of an LPC vocoder with warping, LPCW. This is described in detail in Appendix 5. The reasoning for this type of analysis is as follows. In ordinary linear prediction the speech spectral envelope is modeled by an all-pole spectrum. The error criterion employed guarantees a uniform fit across the whole frequency range. However, we know from speech perception studies that low frequencies are more important than high frequencies for perception. Therefore, a minimally redundant model would strive to achieve a uniform perceptual fit across the spectrum, which means that it should be able to represent low frequencies more accurately than high frequencies. In an attempt to achieve such a uniform perceptual fit, we applied our linear predictive

spectral warping technique to LPC vocoding. The resulting vocoder, denoted by LPCW, can either improve the vocoded speech quality for a given bit rate or lower the bit rate for a given speech quality.

Briefly, at the transmitter of the LPCW vocoder, the short-time speech spectrum is warped such that high frequencies are compressed relative to low frequencies, in the sense that frequency resolution is better at low frequencies than at high frequencies (but spectral amplitudes are not affected by this warping); regular LPC analysis is then performed on the warped spectrum. At the receiver, the all-pole spectrum computed from the decoded parameters is dewarped using the inverse of the warping function, and then regular LPC analysis is carried out on the dewarped spectrum. LPC coefficients resulting from the last step are in turn employed in synthesizing the speech waveform. Synthesis experiments performed using the LPCW vocoder indicated that the introduction of spectral warping produced a saving of about 10-15% in bit rate without affecting the speech quality. The indicated saving, however, is achieved at the expense of increased computation relative to a regular LPC vocoder.

3. PARAMETER QUANTIZATION

The parameters of our LPC vocoder are: log area ratios (LARs), pitch and gain. We developed improved quantization schemes for LARs and pitch. As gain parameter, one can transmit either the energy of the speech signal, or the energy of the prediction error signal. Through statistical error analysis, we determined which of these two energies led, in general, to a smaller quantization error. Details of our work on these issues are given below.

3.1 Quantization of Log Area Ratios

In our previous work we showed that linear or uniform quantization of LARs is optimal in the sense of a minimax spectral error criterion [2]. In deriving this result we used a prototype spectral sensitivity characteristic of the reflection coefficients, which was obtained by averaging spectral sensitivity over a number of speech sounds and over different reflection coefficients. The resulting quantization scheme had the same step size for quantizing all the LARs. However, when we averaged the spectral sensitivity of each reflection coefficient separately over a number of speech sounds, we found that while the sensitivity curves of the different reflection coefficients had the same general U-shape, they were located at different sensitivity levels. By taking advantage of these differences in

sensitivity levels of the reflection coefficients or equivalently LARs, we developed an improved quantization scheme that uses unequal step sizes for the different LARs.

LAR Sensitivity Plots

Employing the experimental procedure that we proposed in our previous work [2], we computed the spectral sensitivity of each LAR and averaged it over a number of speech sounds. Our speech data base consisted of 12 utterances (from 6 males and 6 females) of a total duration of about 30 sec; speech was low-pass filtered at 5 kHz and sampled at 10 kHz. A 12-th order linear prediction analysis was carried out on frames of 20 ms duration of preemphasized speech; we used the first-order preemphasis filter $(1-.969\ z^{-1})$. LPC analysis produces the reflection coefficients $\{k_i\}$, which are related to the LARs $\{g_i\}$ expressed in decibels by the one-to-one mapping [2]:

$$g_i = 10 \log_{10} \frac{1+k_i}{1-k_i}, \quad 1 \le i \le p=12 \tag{3.1}$$

We computed the sensitivity of each of the 12 LARs at 13 equally-spaced points over the range -18 to 18 dB, as follows. The value of, say, the i-th LAR was set equal in turn to one of those 13 values, while the other 11 LARs were kept constant at their respective values obtained through LPC analysis for that frame. g_i was then perturbed by a small amount, and the corresponding change in the spectrum of the linear predictor and

thus the sensitivity of g_i were measured, as explained in our paper [2]. The sensitivity measurement procedure was repeated for each of the other 11 LARs. The 13 sensitivity values of individual LARs were then averaged separately over 25 voiced frames and 15 unvoiced frames, selected from our data base. Figures 3.1 and 3.2 depict averaged spectral sensitivity curves of individual LARs for respectively voiced and unvoiced speech sounds. Each figure has 12 sensitivity curves corresponding to 12 LARs and also an average of all the 12 sensitivity curves. (We have assumed a linear variation in sensitivity between the computed 13 values.)

Average Sensitivity Levels

In order to derive the step sizes for quantizing the LARs, first we need to transform, for each LAR, its sensitivity curve to one number which we shall call its average sensitivity level. For the ith LAR g_i, it is reasonable to define its average sensitivity level S_i as

$$S_i = \sum_{k=1}^{L_i} P_{ik} \left.\frac{\partial S}{\partial g_i}\right|_{g_i=G_{ik}} , \qquad (3.2)$$

where the range of g_i is represented by L_i equally spaced points G_{ik}, $1 \le k \le L_i$; $\partial S/\partial g_i$ is the spectral sensitivity of g_i; P_{ik} is the probability of g_i taking the value G_{ik}. It is clear that S_i is approximately equal to the expected value of $\partial S/\partial g_i$ if L_i is sufficiently large.

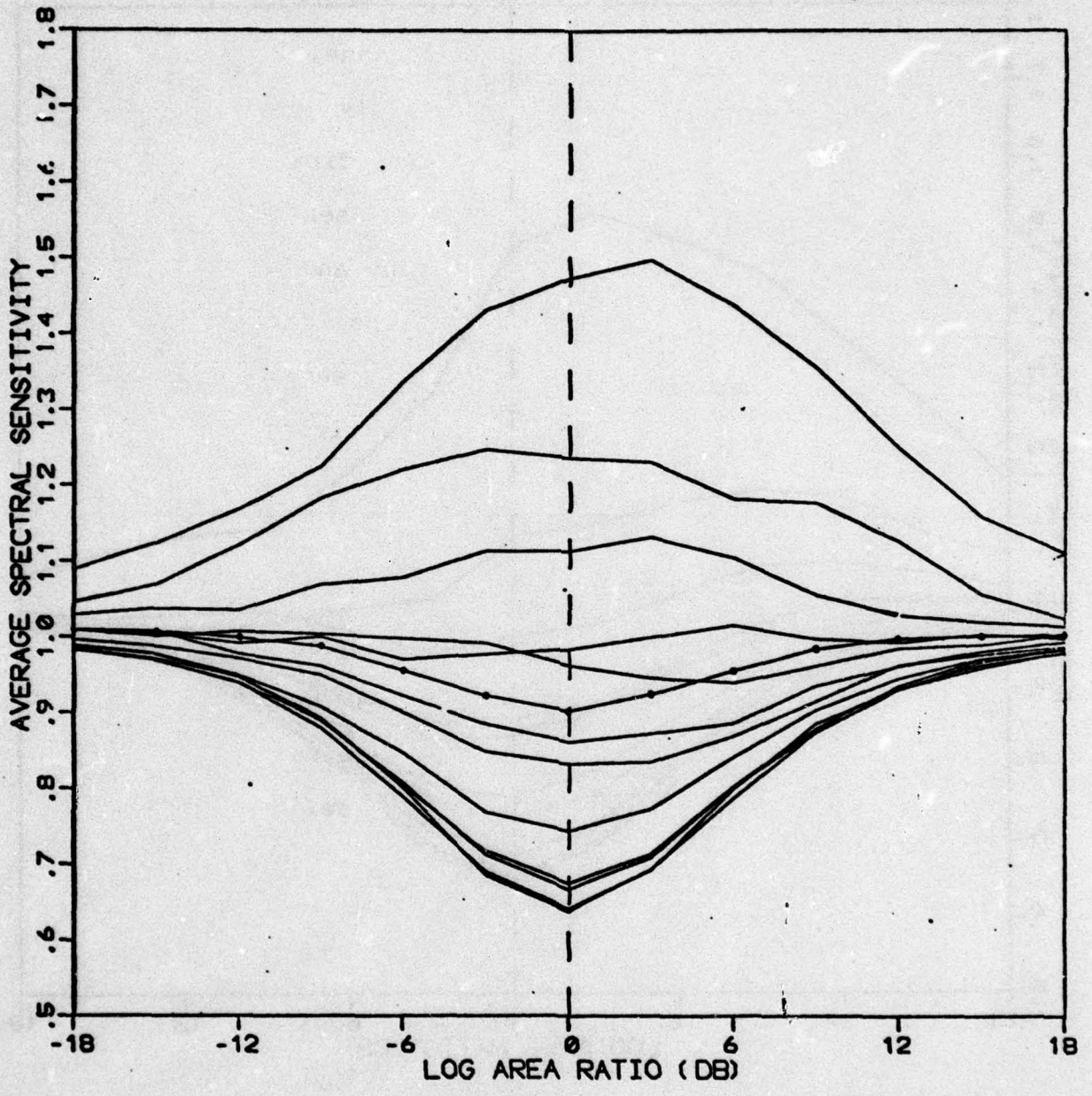


Fig. 3.1 Spectral sensitivity curves for LARs of a 12th order linear predictor, averaged over voiced sounds only. The top curve corresponds to the first LAR; the bottom curve to the 12th LAR. Some sensitivity curves cross each other as shown. The average of the 12 sensitivity curves is drawn along circled points.

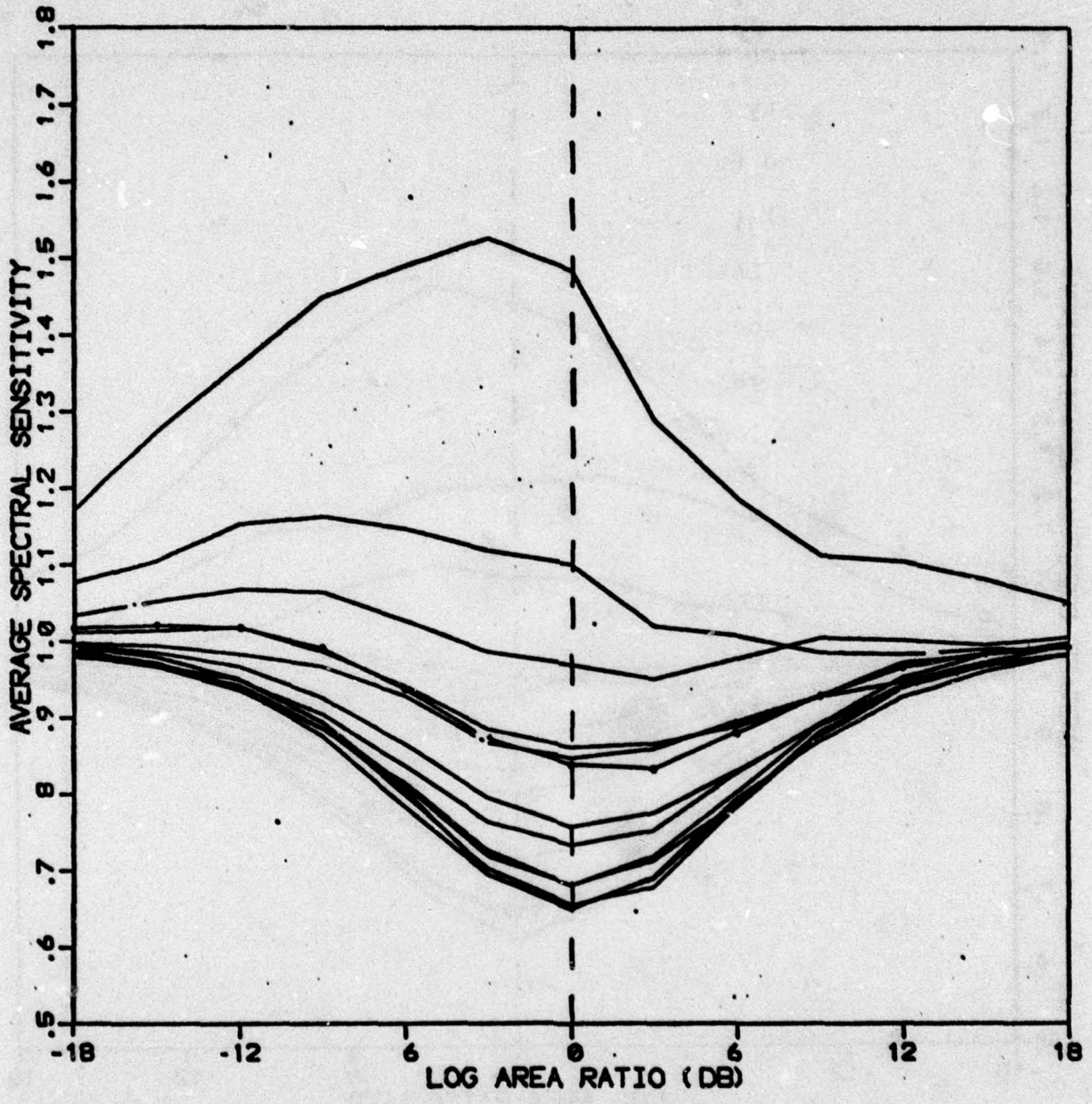


Fig. 3.2 Spectral sensitivity curves for LARs of a 12th order linear predictor, averaged over unvoiced sounds only. The top curve corresponds to the first LAR; the bottom curve to the 12th LAR. Some sensitivity curves cross each other as shown. The average of the 12 sensitivity curves is drawn along circled points.

In computing the quantities S_i for both voiced and unvoiced cases, we used the sensitivity data shown in Figs. 3.1 and 3.2 and the probability histogram data for LARs that we previously collected for Huffman coding purposes [1]. We mention here that those histograms were computed at 1 dB intervals (or bin size) from a 100 frames/sec linear prediction analysis of the preemphasized speech from the above data base. The computed average sensitivity levels S_i, $1 \leq i \leq 11$, are given in Table 3.1, for both voiced and unvoiced cases. Notice that S_i decreases almost monotonically with increasing i and that the sensitivity level of the first LAR is almost twice as much as that of the 11-th LAR. The unequal-step-size quantization method described below takes advantage of this variation in sensitivity levels in determining the various LAR step sizes.

Quantization Method

Using the approach of optimal bit allocation strategy that we presented earlier [2], we computed the number of quantization levels N_i and the step sizes δ_i for the different LARs as follows. The total spectral deviation ΔS due to LAR quantization errors Δg_i, $1 \leq i \leq p$, where p is the LPC order, is given approximately by

$$\Delta S = \sum_{i=1}^{p} S_i |\Delta g_i| \qquad (3.3)$$

LAR #	Average Sensitivity Level	
	Voiced	Unvoiced
1	1.31	1.33
2	1.21	1.03
3	1.11	0.97
4	0.99	0.87
5	0.97	0.86
6	0.87	0.77
7	0.84	0.75
8	0.78	0.70
9	0.71	0.71
10	0.70	0.68
11	0.68	0.67

Table 3.1 Average Sensitivity Levels of Log Area Ratios

In an attempt to minimize the maximum spectral deviation, we replace $|\Delta g_i|$ by its maximum value, which is equal to half the corresponding step size for the linear quantization of g_i using round-off arithmetic. (If truncation arithmetic is used, the maximum value will be twice as much, but the constant scale factor does not change the solution to the minimization problem given below.) Thus

$$(\Delta S)_{max} = \frac{1}{2} \sum_{i=1}^{p} S_i \delta_i \tag{3.4}$$

where

$$\delta_i = [(g_i)_{max} - (g_i)_{min}]/N_i, \tag{3.5}$$

and (g_i)max and (g_i)min are the upper and lower bounds on g_i. The problem is to minimize $(\Delta S)_{max}$ with respect to $\{N_i\}$ subject to the constraint that the total number of bits used for quantizing p LARs be equal to a prespecified value M:

$$\sum_{i=1}^{p} \log_2 N_i = M. \tag{3.6}$$

The solution to the above constrained minimization problem is given below:

$$N_1 = K_1 \left[\frac{2^M}{\prod_{i=1}^{p} K_i} \right]^{1/p}$$

$$N_i = \frac{K_i}{K_1} N_1, \quad 2 \le i \le p, \tag{3.7}$$

where $$K_i = [(g_i)_{max} - (g_i)_{min}]\ S_i,\ 1 \leq i \leq p. \qquad (3.8)$$

To compare unequal step size quantization with equal step size quantization, we have listed in Table 3.2 the numbers of quantization levels for these two methods with the same total number of bits and considering voiced and unvoiced cases separately. As expected, relative to the equal step size method, the unequal step size method places more emphasis on the first three LARs by allotting more levels to them. Synthesis experiments showed that use of the unequal step size quantization method produced better quality speech. The perceived quantization noise in the synthesized speech was reduced noticeably when the transmission rate was very low (e.g., 1800 bps).

It should be noted that for real-time implementation, while the equal step size method requires only one coding table and one decoding table, the unequal step size method in general requires p coding tables and p decoding tables.

3.2 Pitch Quantization

Quantization of pitch presents an altogether different problem from the quantization of other transmission parameters. The major difference is that the decoded pitch values are constrained to be integers (samples per pitch period). Another difficulty arises in attempting to quantize the log pitch in that

Coeff. #	VOICED (43 BITS)		UNVOICED (41 BITS)	
	Equal Step (1dB)	Unequal Step	Equal Step (1db)	Unequal Step
1	28	43	29	51
2	22	31	21	28
3	19	24	14	18
4	15	17	13	15
5	14	15	10	11
6	13	13	9	9
7	13	12	12	11
8	14	12	11	10
9	12	10	10	9
10	11	9	10	9
11	9	7	9	8

Table 3.2 Quantization Levels

at the high frequency end (small pitch period) of the range of interest, the quantization bin size, as found by dividing the log pitch scale into equal segments, can be smaller than the distance between two allowable pitch values (for decoding). This leads to cases where two distinct quantization bins yield the same decoded value, thus wasting some quantization levels. In ARPA NSC Note #49 [3], we proposed a method for deriving the pitch encoding and decoding tables in such a way that maximum usage is made of the different quantization levels. Our simulation system was modified to use this improved pitch quantization scheme. Considering pitch frequencies over the range 50-450 Hz and using 6 bits for quantization, the improved coding/decoding tables are given in Table 3.3. The quantization level 0 denotes unvoiced frame. When the pitch period in number of samples is greater than or equal to C(i), the i-th entry in the column C, but less than C(i+1), then it is coded as level i and decoded as D(i) samples. For example, a pitch period of 100 samples is coded as level 44 and decoded as 101 samples. A pitch period less than 21 samples is coded as level 1, and similarly a pitch greater than 200 samples is coded as level 63.

Statistics of differences in quantized pitch values using the above scheme were collected for a number of speech utterances from male and female speakers for use in Huffman coding of pitch.

C	K	D	C	K	D
21.500			40.505		
	1	22		18	41
22.502			41.990		
	2	23		19	43
23.502			43.517		
	3	24		20	44
24.502			44.978		
	4	25		21	46
25.502			46.558		
	5	26		22	47
26.502			47.819		
	6	27		23	49
27.502			50.175		
	7	28		24	51
28.501			51.495		
	8	29		25	52
29.504			52.867		
	9	30		26	54
30.493			55.043		
	10	31		27	56
31.534			56.984		
	11	32		28	58
32.374			59.038		
	12	33		29	60
33.996			60.879		
	13	35		30	62
35.633			63.487		
	14	36		31	65
36.498			66.178		
	15	37		32	67
37.375			67.829		
	16	38		33	69
39.027			70.532		
	17	40		34	72
40.505			73.045		

Table 3.3 Pitch Coding/Decoding Tables

C	K	D	C	K	D
73.045			118.998		
	35	74		49	121
75.324			123.032		
	36	77		50	125
78.674			126.903		
	37	80		51	129
80.999			131.397		
	38	82		52	134
83.360			136.541		
	39	85		53	139
86.575			141.490		
	40	88		54	144
89.368			146.544		
	41	91		55	149
92.994			151.371		
	42	95		56	154
96.670			157.027		
	43	98		57	160
99.363			162.546		
	44	101		58	165
102.907			167.854		
	45	105		59	171
107.034			174.074		
	46	109		60	177
110.998			179.904		
	47	113		61	183
115.010			186.371		
	48	117		62	190
118.998			193.670		
				63	197
			200.000		

(Table 3.3 continued)

3.3 Gain Quantization

As gain parameter, one can transmit either the energy of the speech signal, $R_\emptyset$, or the energy of the prediction error, E_p. These two quantities are related to each other by:

$$E_p = R_\emptyset \, V_p, \tag{3.9}$$

where V_p denotes the normalized error of the linear predictor. It can be shown that E_p has a smaller dynamic range and hence leads to a smaller quantization error than $R_\emptyset$. However, when transmitting E_p, a problem arises from the fact that the normalized error of the quantized predictor is different from the unquantized case. This causes an error in the energy of the synthesized speech even when E_p is not quantized before transmission. This of course is not the case if we transmit $R_\emptyset$. Another consideration in deciding which transmission parameter to use for gain is the type of synthesizer implementation. Regular filter realization (direct form or ladder structure) and normalized filter realization [4] are the two types used by the NSC group. The gain of the regular filter is equal to the square root of E_p, while the gain of the normalized filter is equal to the square root of $R_\emptyset$. Thus, for example, if the receiver employs the normalized filter, it is better to transmit $R_\emptyset$ since transmitting E_p in this case requires computing the normalized error of the synthesizer filter and dividing with it the received

E_p to obtain the normalized filter gain. Avoiding these extra operations may be desirable particularly for real-time implementation.

We conducted a statistical error analysis using both R_g and E_p for transmission [5]. Our findings indicated that, in general, it is better to use R_g for transmission than to use E_p. Such a choice is more strongly recommended when using the normalized filter. The results of this study also suggested a third alternative which is to transmit the product of R_g and the normalized error of the quantized predictor. This alternative seems attractive for the case when the regular filter realization is used.

4. VARIABLE FRAME RATE TRANSMISSION

4.1 Review of our Past Work

In our previous work on developing minimally redundant narrowband speech transmission systems, we have used quite successfully the concept of variable frame rate (VFR) transmission [1]. In a VFR scheme, model parameters (LPC parameters, log pitch, log gain) are transmitted only when the properties of the speech signal have changed sufficiently since the preceding transmission; the parameters for the untransmitted frames are regenerated at the receiver through linear interpolation between the parameters of the two adjacent transmitted frames. For example, speech parameters may be transmitted less often during steady-state portions of speech, and more often during rapid speech transitions.

Below, we briefly review the particular VFR transmission scheme that we employed in our past work. Linear predictive analysis was performed once every 10 ms on speech, low-pass filtered at 5 kHz and sampled at 10 kHz, to extract 100 frames/sec (fps) of LPC data: pitch, gain and 11 log area ratios (LARs). Pitch and gain were transmitted at the full 100 fps rate, while LARs were transmitted at a variable rate using the following VFR scheme. The transmission scheme computed the distance or the amount of deviation between the LARs of the

current frame and the LARs of the last transmitted frame, and compared this distance against a preselected threshold. The LARs of the current frame were transmitted only when the above distance exceeded the threshold. To compute the above distance, we used the so-called log likelihood ratio measure, which is the logarithm of the ratio of the mean-squared values of the linear prediction error signal obtained for the current frame (i) when the optimal linear predictor parameters (i.e., the LARs extracted for the current frame) are used and (ii) when the last transmitted parameters are used.

During the first year of this contract, we investigated several modifications to the above VFR scheme [6]. An important result of this work is the double-threshold scheme, which compared the log likelihood ratio between a current frame and the previously transmitted frame against two thresholds LRT1 and LRT2, where LRT2>LRT1. If the log likelihood ratio was less than LRT1, the current frame LARs were not transmitted; if it exceeded LRT1, but not LRT2, then the current frame LARs were transmitted; if it exceeded both thresholds, then the LARs of the frame immediately preceding the current frame were transmitted. The purpose of the last step was to avoid having to do parameter interpolation at the receiver between largely different data frames.

The above VFR scheme is being used in the real-time ARPA-LPC System II, whose specifications we provided in the form of an NSC note [7]. This note provides a step-by-step description for both the single-threshold and the double-threshold VFR schemes.

Employing the above VFR scheme, we reduced the LAR transmission rate from 100 fps to an average of about 37 fps, with only a small change in the quality of the resynthesized speech relative to the case when all the available 100 fps data were transmitted. Further, we observed that any significant reduction in the frame rate below 37 fps introduced, in general, noticeable distortions in the speech quality.

In an effort to further reduce the average frame rate of LAR transmission, without speech quality degradation, we developed a new VFR scheme based on a functional perceptual model of speech. The model and the new scheme are described in the next subsection.

4.2 Perceptual-Model-Based VFR Scheme

A detailed description of our perceptual model of speech and manual and automatic VFR schemes based on this model is contained in a paper which is reproduced here as Appendix 6. Below, we briefly review the model and give the details of an improved automatic VFR scheme that we developed since the time the above paper was written.

4.2.1 A Perceptual Model of Speech

With the motivation of developing an efficient VFR transmission scheme, we formulated the following perceptual model of speech:

1) Speech can be represented in terms of LPC (or other) parameters extracted at a minimal set of perceptually significant time points (or frames), not necessarily equally spaced.

2) Between any two such time points, the parameters vary linearly.

3) The location of these points is obtained independently for pitch, gain, and spectral (or LPC) parameters.

Our requirement is that the quality of the resynthesized speech based on this model should be no worse than that of the unreduced or the full 100 fps case. We experimentally demonstrated the validity of the above model by using a manual, trial-and-error scheme, and we achieved a lower limit for the LAR transmission frame rate of about 2 transmissions per phoneme, or about 24 fps. We then developed a fully automatic two-stage scheme which approximately met the model requirements as well as achieved this lower limit of 24 fps (for LAR transmission). Details on the manual and automatic schemes are given in Appendix 6.

A major difference between the perceptual-model-based VFR scheme and our earlier VFR scheme that is being used in ARPA LPC-II system is in the transmission strategy: our earlier scheme performs an "end-to-end comparison," illustrated in Fig. 4.1a, between the preceding transmitted frame and the current frame being considered for transmission; on the other hand, the new scheme as shown in Fig. 4.1b, compares LPC parameters of every frame in the transmission interval with those obtained by linear interpolation between the two "end-frames" and computes the total transmission error as some weighted average of the individual frame errors. It is this difference which has led to a substantially lower transmission frame rate for the new scheme than for our earlier scheme.

Below, we report on several modifications that we made on the two-stage VFR scheme for the transmission of LARs.

4.2.2 Transmission Error Computation

Given that LARs of the frame N, say, have been transmitted, the basic strategy is to determine the longest line extending from $\underline{g}(N)$ (vector of p LARs for frame N) in the p-dimensional parameter space such that the resulting transmission error computed between the actual parameter vectors $\underline{g}(N+i)$ and the interpolated parameter vectors $\hat{\underline{g}}(N+i)$ over the duration of that line is less than some threshold (see Fig. 4.1b). First, we need

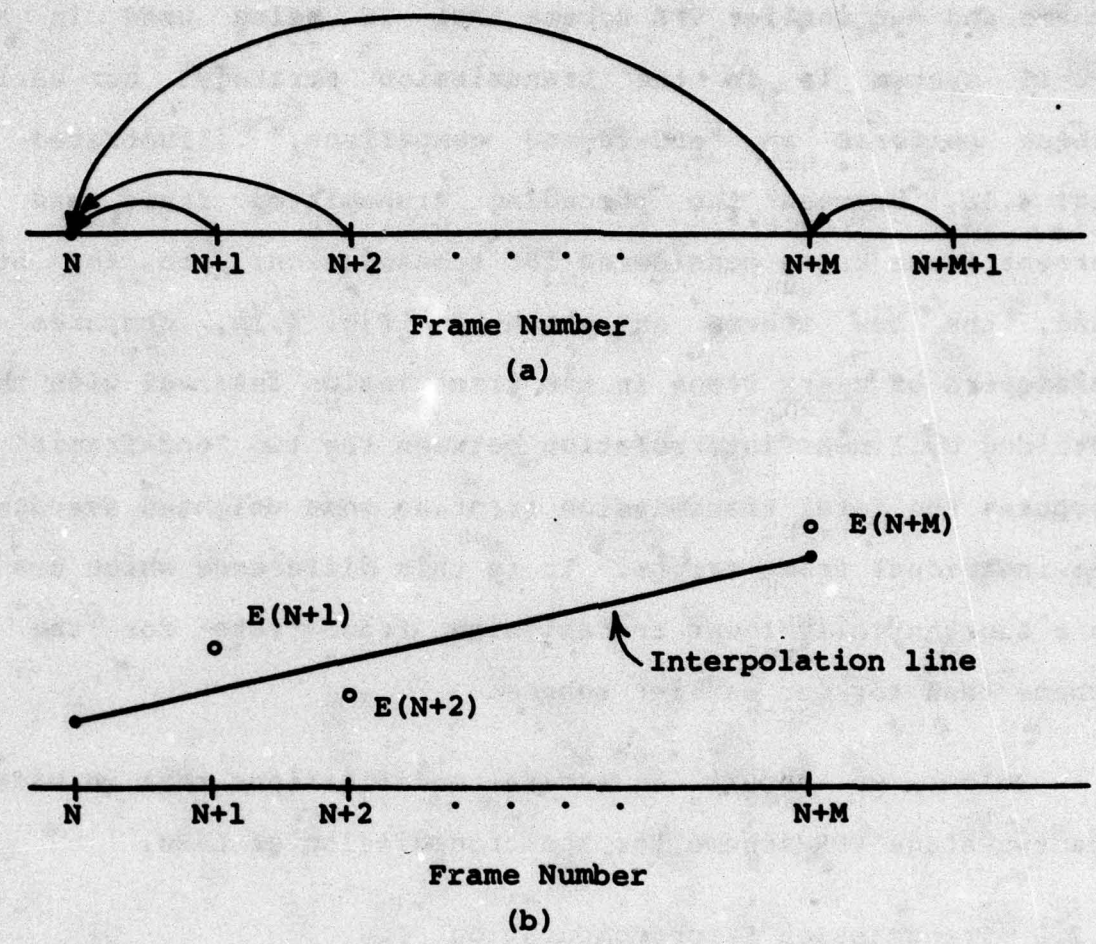


Fig. 4.1 Illustration of two VFR transmission strategies.

(a) "End-to-end" error measurement of our old VFR scheme used in LPC-II.

(b) Average frame error between actual and interpolated values computed over the transmission interval. E(N+i) is the error for frame N+i. The frame error E(N+M) is due to parameter quantization (see Section 4.2.3).

to define frame error, or distance between two sets of LARs $\underline{g}$ and $\hat{\underline{g}}$ for any given frame, and then specify how this error is averaged over several frames (time averaging).

Frame Error

The frame error for frame n, denoted by E(n), is defined as the weighted Euclidean distance:

$$E(n) = \sum_{i=1}^{m} w_i [g_i(n) - \hat{g}_i(n)]^2 / \sum_{i=1}^{m} w_i , \qquad (4.1)$$

where $\{w_i\}$ is the set of coefficient weights chosen to reflect the relative importance of the different LARs (presumably to perceived speech quality); we allow $m \leq p$.

We have chosen the coefficient weights to be the expected or average spectral sensitivities of individual LARs (see Table 3.1). For the first 4 LARs, these are: 1.3, 1.2, 1.1 and 1.0. This weighting scheme is based on the reasonable idea that a given amount of error in a LAR with a higher sensitivity is more important to spectral accuracy (and hence perception) than the same error in a LAR with a lower sensitivity. Surprisingly, however, our experimental results showed that different choices of these weights (e.g., w_i all equal to 1) produced no discernible change in speech quality.

We found through experimentation that the summation in the frame error definition (4.1) need be done only up to the first 4 LARs (i.e., m=4).

Another way to compute the frame error is via log likelihood ratio measure explained above. Our experiments (see Subsection 4.2.7) indicated identical speech quality results for the same average transmission frame rate, for the two measures: LAR distance and log likelihood ratio. Since LARs are being used as transmission parameters, use of the LAR distance measure is computationally much less expensive than the log likelihood ratio measure. So, we employed the LAR distance in all our subsequent experiments.

Transmission Error

The transmission error ET between frames N and N+M is computed as the weighted, time-averaged frame error:

$$ET = \frac{1}{M} \sum_{n=N+1}^{N+M} W(n)E(n), \qquad (4.2)$$

where W(n) is the frame weight for frame n. (The upper limit for the summation in (4.2) is considered as N+M to incorporate the effect of LAR quantization; E(N+M) is computed from (4.1) with $\hat{g}_i$ denoting quantized LAR values.) As frame weight, we successfully used the speech signal energy per sample in that frame, RØ, expressed in decibels and normalized with respect to some estimate RM of the maximum value of RØ:

$$W(n) = R\emptyset(n)/RM(n). \qquad (4.3)$$

The idea behind the weighting scheme given by (4.3) is that even large frame errors do not make a perceptible effect if they are associated with relatively small speech signal energies. For our speech data base, where we have 9-bit samples, RØ is usually around 35-4Ø dB for open vowels, 15-3Ø dB for fricatives, and around Ø-7 dB for the silent period of an unvoiced plosive.

A simple and efficient way to update RM is by the following recursive method:

$$RM(n) = \text{Max}\{R0(n), \alpha\, RM(n-1), 25\ dB\}, \qquad (4.4)$$

where α is a constant less than 1. We use α =Ø.98, which means that RM decays to half its original value in about 27 frames. It should be noted from (4.3) and (4.4) that W(n)=1 if RØ(n)>25 and has been increasing or has been decreasing slowly; W(n)<1 if RØ(n)<25 or if RØ(t) has been decreasing at a faster rate than exp(-Ø.98t).

4.2.3 Parameter Quantization

There are two ways in which the effect of parameter quantization can be included within the above procedure for transmission error computation. Both ways can be employed simultaneously.

First, since the transmitted LARs have to be quantized, we consider the interpolation line between the quantized LARs of the two end-frames (frames N and N+M in (4.2)). A frame error is then computed as the distance given by (4.1) between the unquantized LARs of that frame and the corresponding LARs obtained from the above interpolation line. The frame error for the right end-frame (E(N+M) in (4.2)) is entirely due to parameter quantization.

The second way of incorporating parameter quantization is what we call the "adjustable" quantization method. A parameter value is normally quantized to its nearest quantization level. The adjustable quantization scheme allows either of the two nearest quantization levels. Thus, given the quantized LARs of the initial frame (left end-frame), the scheme determines the adjusted quantized values of the LARs for the final frame (right end-frame) in the transmission interval, in such a way that the total transmission error is minimized.

A one-dimensional (p=m=1) example is shown in Fig. 4.2 to illustrate the "adjustable" quantization. For this example, the parameter value of the sixth frame is selected for transmission. If this value is quantized to the nearest quantizer output (the output just below it), there is considerable interpolation error in the interval between frames 1

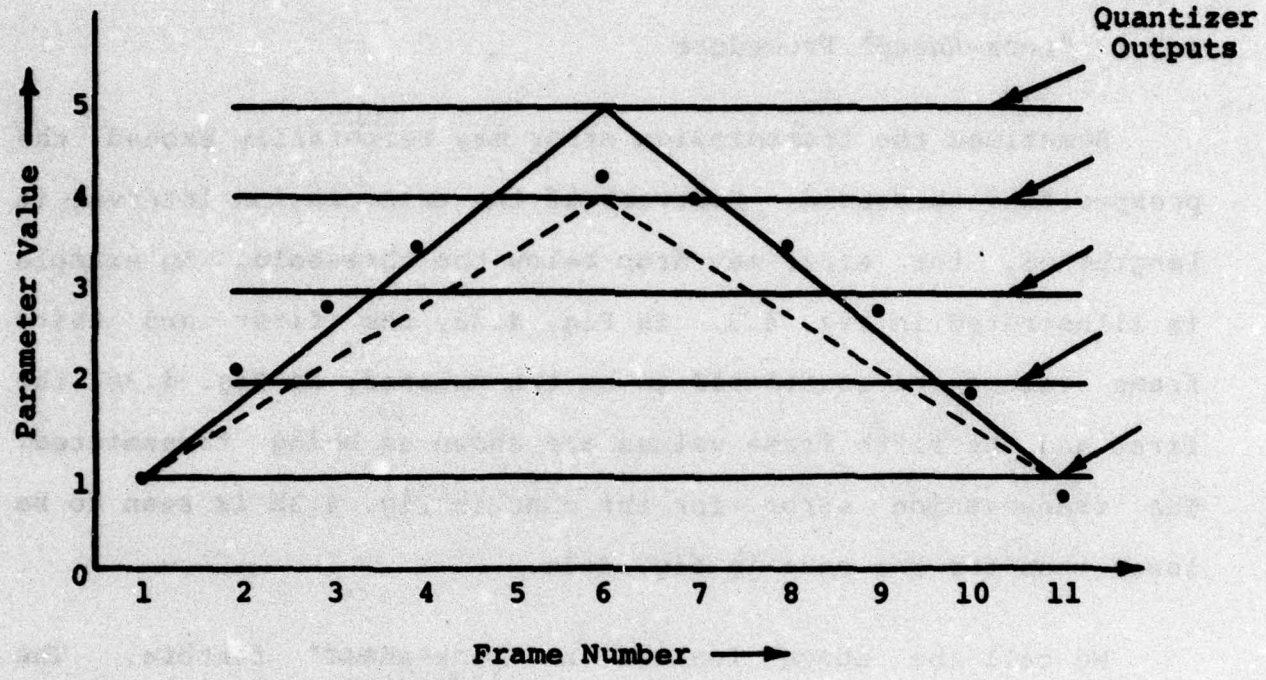


Fig. 4.2 Example to illustrate the "adjustable" quantization scheme. Dashed-line plot corresponds to normal quantization, where a parameter value is quantized to the nearest quantizer output. Solid line corresponds to the "adjustable" quantization (see text). (The dots represent the original unquantized parameter data.)

and 6. If the higher quantizer output is used instead, the total transmission error is reduced. (Fig. 4.2 also shows the interpolation line for the next transmission interval from frame 6 to frame 11.)

4.2.4 "Look-Ahead" Procedure

Sometimes the transmission error may temporarily exceed the prespecified threshold. However, if the transmission interval is lengthened, the error may drop below the threshold. An example is illustrated in Fig. 4.3. In Fig. 4.3a, the first and third frame values are considered to be transmitted; in Fig. 4.3b, the first and the fifth frame values are shown as being transmitted. The transmission error for the case in Fig. 4.3b is seen to be lower than for the case in Fig. 4.3a.

We call the above feature a "look-ahead" feature. The extent of "look-ahead" (in terms of number of frames to consider) is limited only by the resulting computational burden; we use a four-frame "look-ahead" procedure. If the error does not drop below the threshold even after moving forward by four frames, we hypothesize the transmission of the frame immediately preceding the one where the threshold was first exceeded.

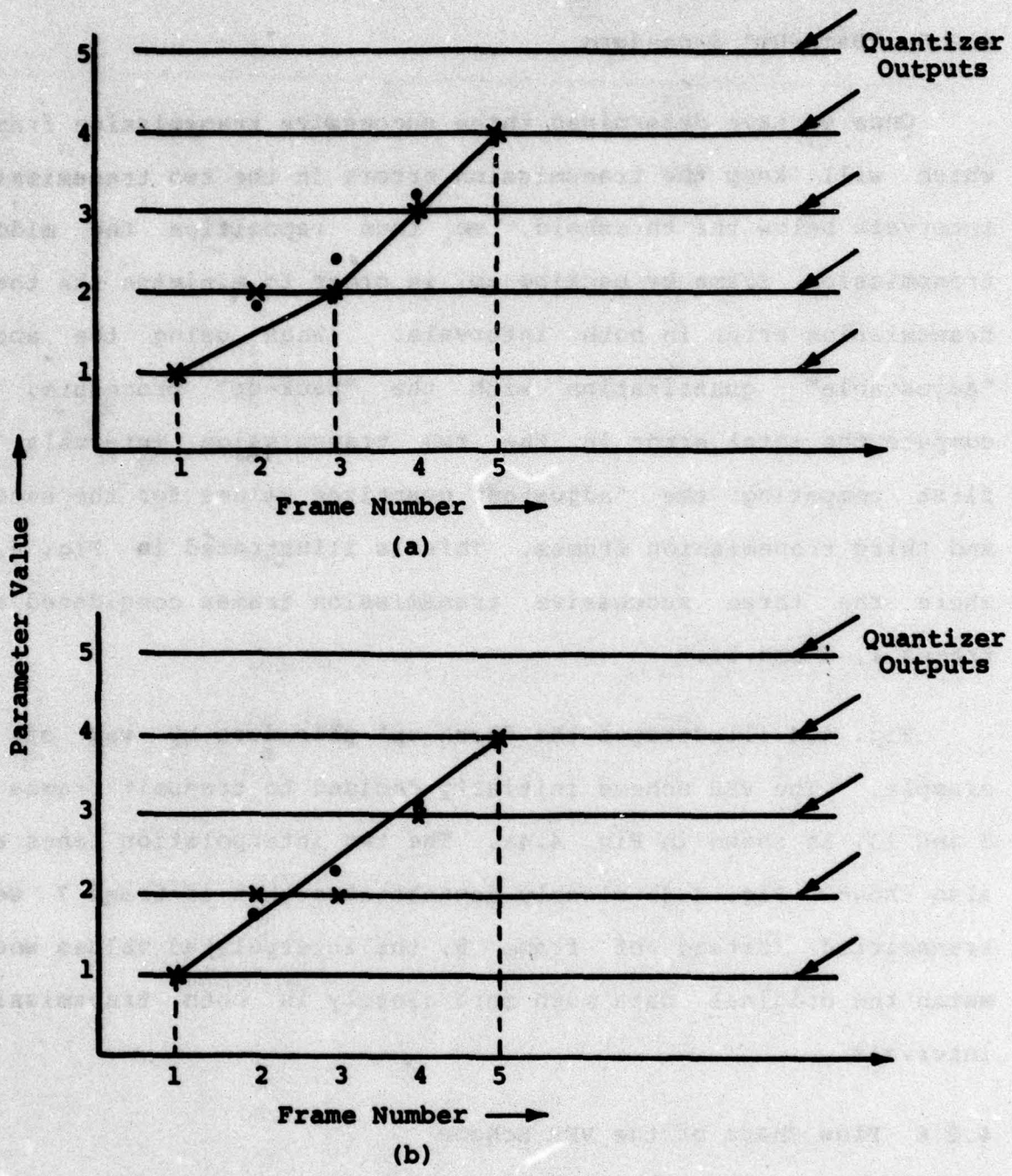


Fig. 4.3 Example to illustrate "look-ahead" procedure. The dots represent the original unquantized parameter values. The x's represent quantizer output values. The vertical dashed lines indicate frames chosen for transmission.
(a) Without the "look-ahead" scheme
(b) With the "look-ahead" scheme

4.2.5 "Back-Up" Procedure

Once we have determined three successive transmission frames which will keep the transmission errors in the two transmission intervals below the threshold, we then reposition the middle transmission frame by backing up, in order to minimize the total transmission error in both intervals. (When using the above "adjustable" quantization with the "back-up" procedure, we compute the total error in the two transmission intervals by first computing the "adjusted" quantized values for the second and third transmission frames. This is illustrated in Fig. 4.2, where the three successive transmission frames considered are frames 1, 6 and 11.)

Fig. 4.4 illustrates the "back-up" procedure by way of an example. The VFR scheme initially decided to transmit frames 3, 8 and 13, as shown in Fig. 4.4a. The two interpolation lines are also shown. Fig. 4.4b clearly demonstrates that if frame 7 were transmitted instead of frame 8, the interpolated values would match the original data much more closely in both transmission intervals.

4.2.6 Flow Chart of the VFR Scheme

The flow chart of the VFR scheme described in the previous subsections is given in Fig. 4.5. Variables that appear in the flow chart are defined in Table 4.1. A function called ERROR is

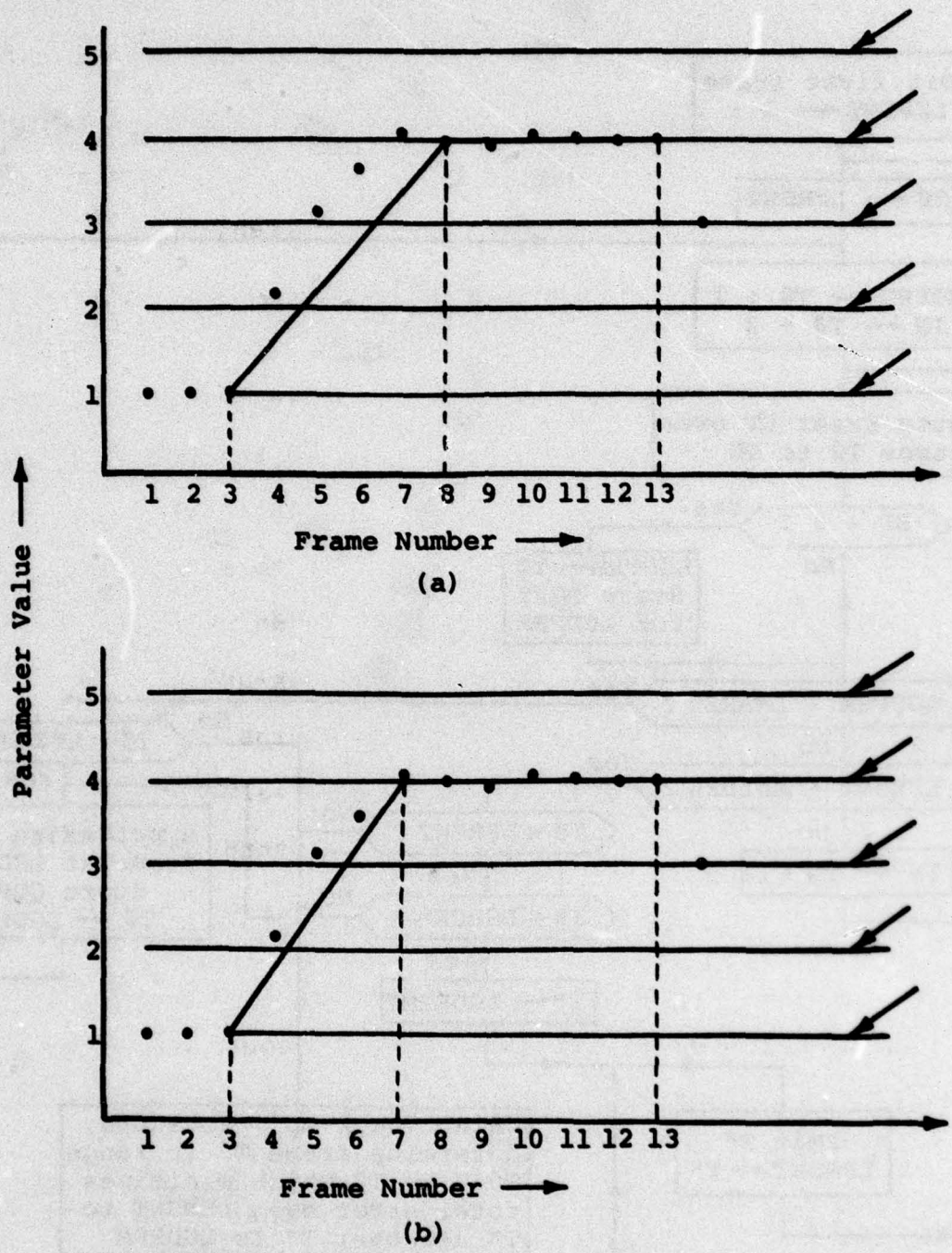


Fig. 4.4 Example to illustrate the "back-up" procedure. The dots represent the original unquantized parameter values. The vertical dashed lines indicate frames chosen for transmission.
(a) Without the "back-up" scheme
(b) With the "back-up" scheme

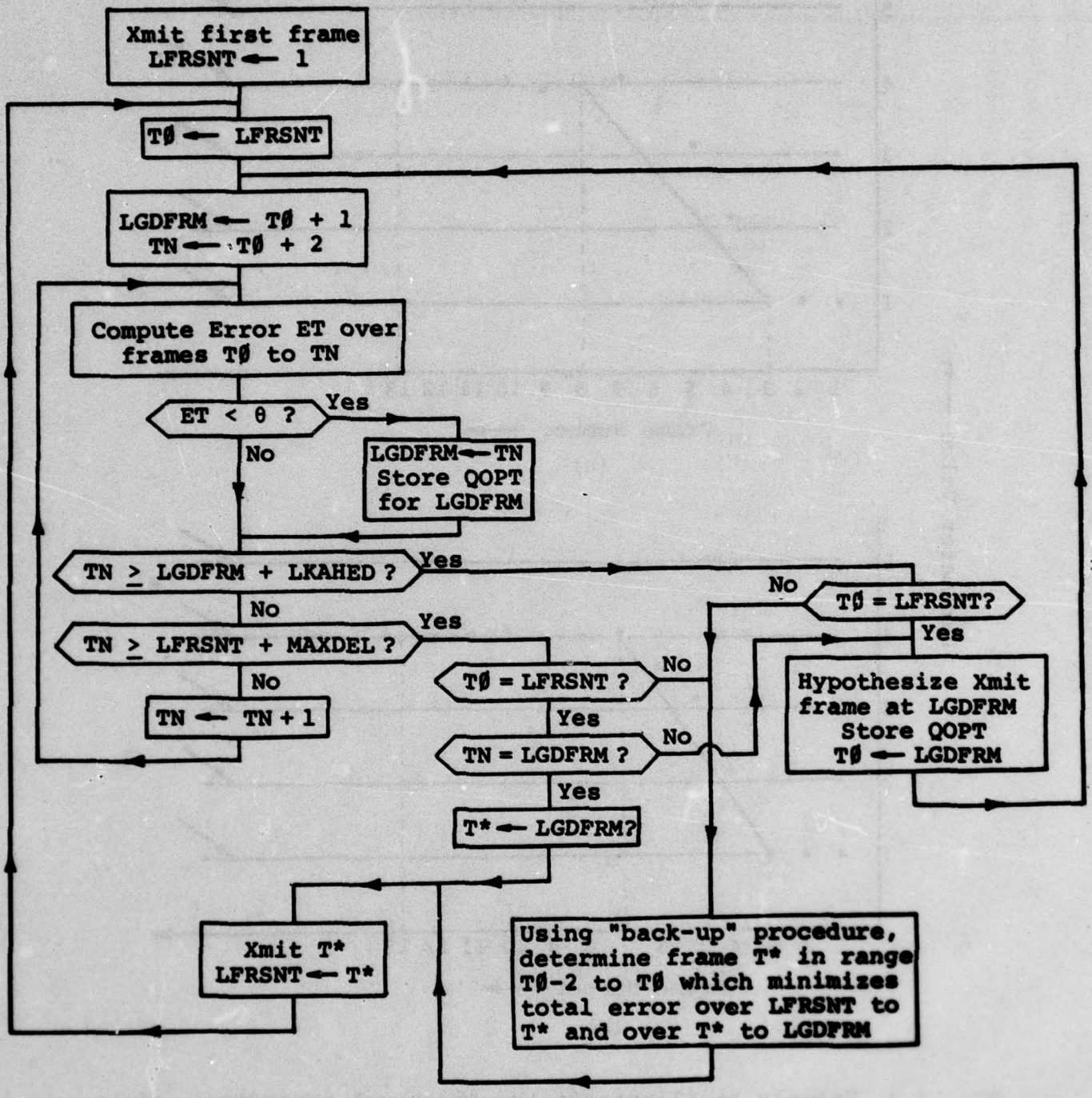


Fig. 4.5 Flow chart of full algorithm.

Table 4.1

List of Variables Used in the Flow Chart in Fig. 4.5.

LFRSNT	Last frame actually transmitted.
TØ	First frame (left end-frame of the interpolation line) in a transmission interval. TØ equals either LFRSNT or a hypothesized transmission frame when using the "back-up" scheme.
TN	Current frame (right end-frame of the interpolation line).
ET	Transmission error between the original unquantized LAR data, and the quantized and interpolated values, computed over the interval from frame TØ + 1 to frame TN (see eq.(4.2) in the text).
θ	Transmission error threshold. Normally, θ=1.3
LGDFRM	Last good frame, i.e., last frame where ET < θ.
LKAHED	Number of frames to "look ahead" beyond the frame where ET exceeds θ. Normally, LKAHED = 4 frames.
MAXDEL	Maximum allowed transmission delay. Without the "back-up" scheme, it is the maximum transmission interval permitted. With the "back-up" scheme, it is the maximum allowed interval between a transmitted frame (LFRSNT) and a frame which is the second hypothesized transmission frame after LFRSNT if it is not LGDFRM, or the second hypothesized transmission frame plus LKAHED, otherwise. Normally, MAXDEL = 12 frames (12Ø ms).
QOPT	Quantized LAR values resulting from the "adjustable" quantization scheme.
T*	Frame position, determined by the "back-up" procedure.

used to compute the transmission error between two hypothesized transmission frames. It accepts as input, quantization levels for the LARs at the first or initial frame, and determines the "adjusted" set of quantization levels for the second transmission frame. If the function is called with three transmission frames, it provides the optimal set of quantization levels for the second and third transmission frames. Each box shown in the flow chart translates into one or two FORTRAN statements.

4.2.7 Experimental Results

We tested the VFR algorithm described above on a set of nine sentences (JB1, DD2, RS3, AR4, DK4, JB5, RS6, DK6 and DD6; 6 sentences from 3 males and 3 sentences from 2 females) from the data base used in our speech quality evaluation work (see Section 7.2.1). Table 4.2 describes six vocoder systems and lists their average transmission frame rates and bit rates obtained over the nine sentences. We ran informal, pair-wise speech quality comparison tests on the syntheses from these six vocoder systems, to evaluate the relative performance of the different versions of the above VFR scheme.

Vocoders 1 and 2 given in Table 4.2 employed the full 100 fps fixed-rate transmission for all parameters (pitch,gain and LARs). Vocoder 1 used the unquantized parameters for synthesis, while Vocoder 2 quantized the parameters using 5 bits for gain, 6

Vocoder System	Parameter Quantization	Adjustable Quantization	Look Ahead	Backup	Pitch & Gain Fixed or Variable Frame Rate	LAR Frame Rate (fps)	Average Bit Rate (bps)
1	No	-	-	-	Fixed	100 (Fixed)	-
2	Yes	No	No	No	Fixed	100 (Fixed)	5650
3	Yes	No	No	No	Variable	30	1850
4	Yes	No	Yes	No	Variable	27	1750
5	Yes	No	Yes	Yes	Variable	27	1750
6	Yes	Yes	Yes	Yes	Variable	25	1650

Table 4.2. Description of six vocoder systems tested and their average transmission frame and bit rates.

bits for pitch (plus 1 bit for Voiced/Unvoiced status), and 44 bits for LARs of voiced frames and 42 bits for LARs of unvoiced frames, which resulted in a transmission bit rate of about 5650 bps. Vocoders 3-6 quantized the parameters in the same way, but employed VFR transmission for all parameters. For pitch and gain VFR transmission, they all used the double-threshold FIT scheme on the quantized values (levels) (see Section 4.3.2), with thresholds of 0 and 1 for pitch, and 1 and 2 for gain; this yielded a transmission frame rate of about 28 fps for pitch and 24 fps for gain. The VFR scheme used for LAR transmission became progressively complex going from Vocoder 3 to Vocoder 6, with Vocoder 6 employing the complete VFR scheme described in the last subsection via flow chart. The simplest VFR scheme (used by Vocoder 3), employs the quantized LARs of the end-frames of the interpolation line (see Subsection 4.2.3). For all the four vocoders, the threshold θ (see flow chart in Fig. 4.5) for the transmission error ET was chosen as 1.3. (We chose m=4 in the expression (4.1) for frame error since it yielded the same speech quality as any higher value but at the least computational effort.) The above choice of the transmission error threshold produced an average frame rate of about 25 fps for the full scheme (Vocoder 6) and an average transmission error (ET averaged over the nine sentences) of 0.55.

Informal tests of pair-wise speech quality comparisons were run for the six vocoders. Also, we compared the full VFR scheme (Vocoder 6) with our earlier "end-to-end" scheme used in LPC-II and with the 50 fps fixed-rate scheme used in LPC-I. (The latter two vocoders we considered were not LPC-II and LPC-I in view of the differences in vocoder conditions such as speech signal sampling rate, bit allocation for parameter quantization, and pitch extraction scheme.) Below, we describe the results of only the important comparisons. (Speech quality tests comparing Vocoders 3-5 with Vocoder 6 are given in Subsection 4.2.8.)

1. Vocoder 2 vs Vocoder 6. There were cases for which speech transitions were more "crisp" for Vocoder 2 (5650 bps) than for Vocoder 6 (1650 bps). However, for most sentences (especially the slowly varying ones, JB1 and DD2), the synthesized speech from Vocoder 2 sounded worse in that it had appreciably more "wobble" quality than the synthesis from Vocoder 6. Our explanation for the observed quality difference is that for the cases when the "wobble" quality is perceived, the error due to parameter quantization is more than the error due to parameter interpolation.

2. Same comparison as in (1), except that both systems used unquantized parameters in the synthesis (i.e., Vocoder 1 vs unquantized version of Vocoder 6). The syntheses for the slowly varying sentences JB1 and DD2 from the variable rate

system had slightly less "wobble" quality than those from the fixed rate system. This is probably due to the fact that small inaccuracies in the LPC analysis arising from interaction between the pitch period and the analysis interval tend to be reduced by the smoothing effect of the interpolation employed by the VFR scheme. There were a couple of situations (during the part "trouble with" in the sentence DK6) where the fixed rate synthesis sounded better. In general, Vocoder 1 and the unquantized version of Vocoder 6 produced speech with essentially the same quality.

3. Vocoder 1 vs Vocoder 6. Surprisingly, the results of this comparison between the unquantized 100 fps system and the 1650 bps VFR system were the same as given above in (2).

4. Vocoder 6 (1650 bps) produced speech quality equal to or better than that of the VFR system with the earlier "end-to-end" scheme of LPC-II (2100 bps). Speech quality improvements observed in the syntheses from Vocoder 6 included clarity and "crispness" of several syllables which were slurred when processed through the earlier VFR system.

5. Vocoder 6 (1650 bps) was compared against the 50 fps fixed-rate system (2825 bps). LPC-I also uses the 50 fps fixed-rate transmission but operates at even a higher bit rate of about 3500 bps. Although the 50 fps system had less

"wobble" quality than the 100 fps system (Vocoder 2), it still had a more "wobble" quality than Vocoder 6, especially for the sentences JB1 and DD2.

6. Finally, we employed the log likelihood ratio measure for computing the frame error between the two sets of LARs $\underline{g}$ and $\hat{\underline{g}}$, instead of the weighted Euclidean distance measure given by (4.1). (Notice that for likelihood ratio computation, LARs are to be first transformed to predictor coefficients.) We adjusted the transmission error threshold (θ) so as to obtain about the same average frame rate (25 fps) as Vocoder 6. We found that the speech quality of the resulting vocoder was identical to that of Vocoder 6. This result leads to the following two observations. First, we conclude that the superior performance of the new perceptual-model-based VFR scheme over the earlier, "end-to-end" scheme of LPC-II (see (4) above), is not due to the change in the definition of the frame error, but due to the difference in the way the transmission error is computed in each case (see Fig. 4.1 which illustrates this difference). Secondly, we recommend the use of the LAR distance measure (4.1) in preference to the log likelihood ratio measure, since the use of the latter measure requires about 50 times more computational time.

4.2.8 Simplified VFR Scheme

Though the algorithm described above produced very low frame rates and good quality speech, it has the disadvantage of being fairly complex, and somewhat slower than real time in our simulation on a KL-10 computer. Of course it could be coded to run in real time on a fast mini-computer, but might not leave enough time for other processing needs. Therefore, we tried several simplifications of the algorithm, in order to arrive at a reasonable compromise between speed, complexity, frame rate (and bit rate) and speech quality.

Our first simplification (see Table 4.2, Vocoder 5) involved the adjustable quantization. Instead of allowing two possible quantization levels for each LAR of every hypothesized transmission frame, the LAR values were always quantized to the nearest levels. This sped up the algorithm by a factor of 4, and reduced the complexity. The transmission frame rate (for the same transmission error threshold) rose to about 27 fps. However the resulting sentences were indistinguishable from those produced by the scheme with adjustable quantization.

For the second simplification we eliminated the "back-up" procedure (Vocoder 4). The frame rate remained unchanged at 27 fps, but the average measured transmission error increased by about 20%. Careful, repeated listening through headphones

revealed only a slight degradation for two sentences. The differences were not perceived through high quality loudspeakers, and were not noticed on single paired-comparisons through headphones. This simplification sped up the algorithm by another factor of 3, and reduced the complexity considerably.

The third simplification was the removal of the "look-ahead" procedure (Vocoder 3). That is, as soon as the transmission error computed over the interval from the preceding transmitted frame to the current frame exceeded the threshold, the frame immediately preceding the current one was chosen to be transmitted. As expected, this increased the frame rate substantially (to 30 fps), for the same speech quality. When the "look-ahead" procedure enabled the algorithm to skip over a bad region, the transmission intervals were greatly lengthened. The simplification reduced processing time by about 30%, and eliminated only 3 lines of FORTRAN code.

Recommended Scheme

While the full scheme (Vocoder 6) clearly results in a lower frame rate and slightly better speech quality, it is much more complex and an order of magnitude slower than the simplest scheme (without "adjustable" quantization, and "back-up" and "look-ahead" features). The first two simplifications discussed above seem reasonable, since the resulting loss was small. The

last feature ("look-ahead") is recommended, since its removal resulted in substantial losses and produced only minor gains.

Fig. 4.6 shows a flow chart of the recommended VFR scheme (Vocoder 4). Comparison of this simplified scheme with Fig. 4.5 will make the difference in complexity apparent.

Of course, if the computer running the VFR algorithm is fast enough, and easy to program, it may be worth the extra trouble to implement the full scheme, which includes the features of "adjustable" quantization and "back-up".

4.3 Transmission of Pitch and Gain

We have developed two types of VFR schemes for the transmission of pitch and gain. These are; (1) "Floating-Aperture Predictor," which performs an "end-to-end" comparison between the parameter values of the current frame and the last transmitted frame, and (2) "Fan Interpolation Technique", which explicitly takes advantage of the fact that the receiver performs linear interpolation for the reconstruction of untransmitted data. The results of our investigation on these two types of schemes are given below.

4.3.1 Floating Aperture Predictor (FAP)

VFR transmission schemes of the FAP type have been described in detail in our NSC Note No. 96 [8]. We developed both

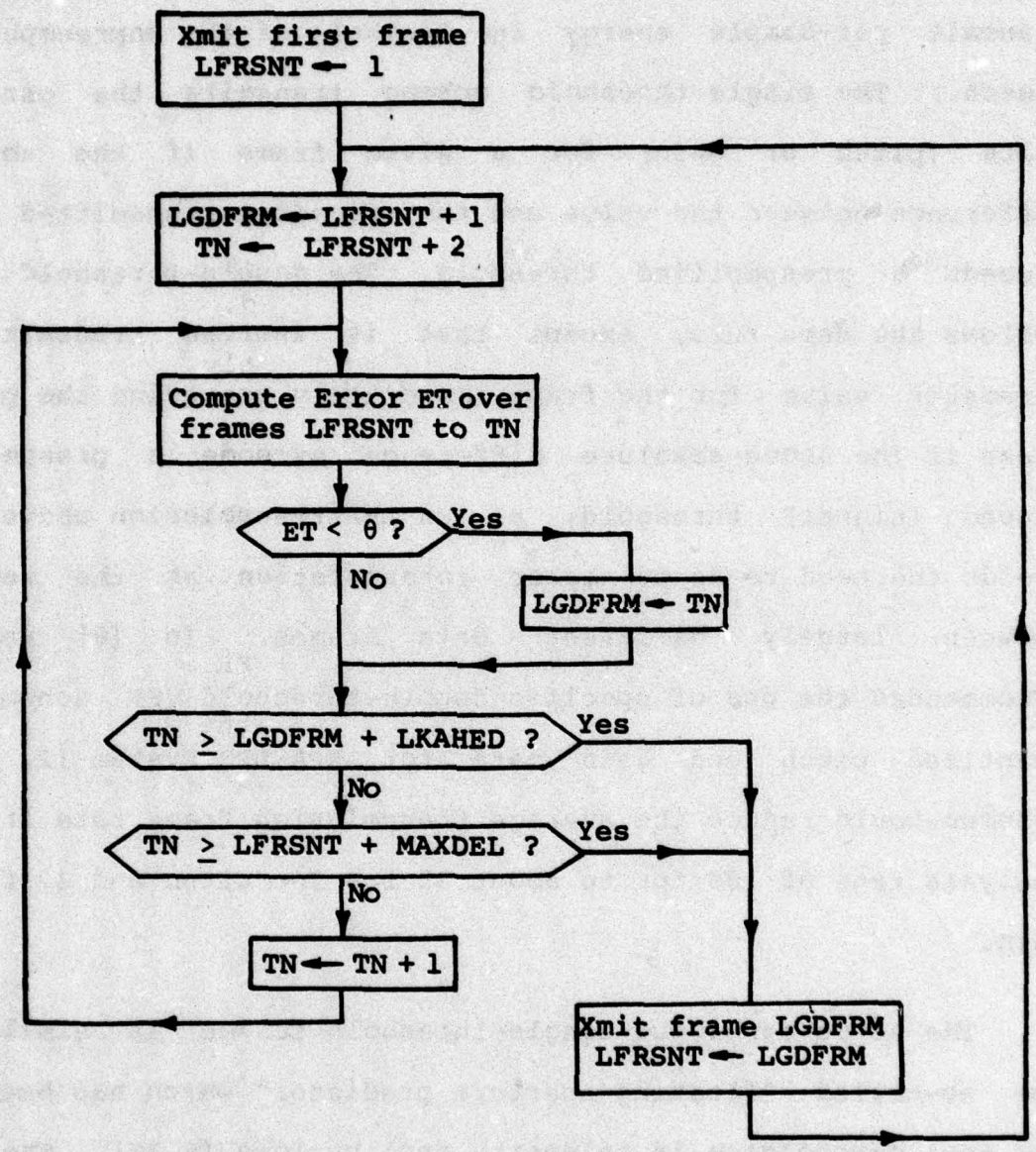


Fig. 4.6 Flow chart of recommended, simplified algorithm.

single-threshold and double-threshold VFR schemes for the transmission of pitch and gain. (As LPC gain parameter, we transmit per-sample energy in decibels of the unpreemphasized speech.) The single-threshold scheme transmits the parameter value (pitch or gain) for a given frame if the absolute difference between the value and the preceding transmitted value exceeds a prespecified threshold. The double-threshold scheme follows the same rule, except that it instead transmits the parameter value for the frame immediately preceding the present frame if the above absolute difference exceeds a prespecified second (higher) threshold; as in LAR transmission above, this avoids the need to do parameter interpolation at the receiver between largely different data frames. In [8] we have recommended the use of specific double-threshold VFR schemes on quantized pitch and gain data for ARPA-LPC System II. These schemes would reduce the average transmission frame rate from the analysis rate of 100 fps to about 35 fps for pitch and 32 fps for gain.

The above-mentioned single-threshold scheme is similar to the so-called "floating-aperture predictor" which has been used for data compression in telemetry applications [9,10]. The main difference between the two is in the way data reconstruction takes place at the receiver i.e., how the untransmitted parameter values are approximated. The traditional FAP method employs a

stair-step reconstruction in that a transmitted value is held constant for all the frames up to the next transmission, where the value is instantaneously updated to be the next-transmitted value. Our single-threshold scheme, however, performs linear interpolation between adjacent transmitted values to generate a smoother approximation. (The double-threshold scheme has the same feature, except that, as mentioned above, it produces less interpolation error at the expense of a slight increase in frame rate.) It is felt that in speech resynthesis applications the smooth approximation produced by interpolation should produce less speech quality distortion (e.g., "roughness") than the stair-step approximation used in the FAP method. However, at the transmitter, our VFR scheme (hereafter loosely called as FAP scheme) does not explicitly take advantage of the fact that the receiver performs linear interpolation for data reconstruction. The inclusion of this feature may perhaps yield further data compression. To this end, we have adapted the so-called "fan interpolation" technique that has been used once again in telemetry applications [9,10].

4.3.2 Fan Interpolation Technique (FIT)

Single-Threshold Scheme: The FIT method previously used in the literature [9,10] is indeed a single-threshold scheme. The method relies on the approximation of the analysis or source data by straight line segments and transmits only those parameter

values corresponding to the end frames of these segments. Given some initial transmitted frame, it finds the longest line for which the maximum error magnitude between the line and the data over the length of the line is below a given threshold. We treated the case where quantized parameter values (levels) are used for deciding when to transmit. In computing the error between the quantized parameter level for a frame and the interpolation line, we compute the interpolated value for that frame, round it off to the nearest (integer) level and then find the difference between this and the actual quantized parameter level for that frame. (Rounding is done such that if the fractional part of the interpolated value is equal to or greater than 0.5 then it is rounded up, otherwise it is rounded down.) At the receiver, quantized levels for untransmitted frames are generated by interpolating between the adjacent transmitted levels and rounding off the interpolated value to the nearest level as explained above.

A step-by-step description of the FIT single-threshold scheme is given in Fig. 4.7 below, where I_n denotes the quantized level of the parameter for frame n, the symbol [ ] refers to the above rounding operation, and T is the preselected threshold.

(1) Transmit value at frame n

$m \leftarrow 2$

(2) $k \leftarrow 1$

(3) $P \leftarrow (m-k)/m\ I_n + k/m\ I_{n+m}$

$E \leftarrow |[P] - I_{n+k}|$

If $E \leq T$, go to (4)

$n \leftarrow n+m-1$

Go to (1)

(4) $k \leftarrow k+1$

If $k \leq m-1$, go to (3)

(5) (No transmission)

$m \leftarrow m+1$

Go to (2)

Fig. 4.7 Description of our FIT single-threshold scheme

It is clear from step (3) that with frames n and (n+m) as end frames, the scheme looks at the magnitude of the interpolation error, in order, from frame (n+1) to (n+m-1) and decides to transmit frame (n+m-1) value at the first instance the error magnitude exceeds T.

If T=0, it is easily seen that the receiver has the same parameter data as at the output of the quantizer. The same result is also achieved using the FAP method with a zero

threshold and with stair-step reconstruction at the receiver. Average transmission frame rates produced by the two methods can, however, be different; the extent of this difference depends upon the nature of the data, in this case quantized parameter levels. For instance, if the data has frequent occurrence of sequences of equal levels (i.e., presence of horizontal or level lines), then the FAP scheme would generally do better yielding a lower frame rate than the FIT method; the reason for this is that the latter method transmits both end frames for each level line, while the former transmits only the first end frame. On the other hand, if the data involves a large number of sloped or nonlevel lines then the opposite result is true in that the FIT method yields a lower frame rate.

Experimental results obtained using the above FIT method on quantized pitch and gain are reported in the sequel.

Double-Threshold Scheme: The double-threshold version of the FIT method operates as follows. Assume that frames n and (n+m) are the end frames of the interpolation line under consideration. Then, (1) if the maximum interpolation error magnitude over the length of the line exceeds the second (higher) threshold T2, then frame (n+m-1) value is transmitted; (2) if the maximum error magnitude exceeds the first (lower) threshold T1, and not T2, then frame (n+m) value is transmitted; (3) if the maximum error magnitude does not exceed T1, then a new interpolation line is

considered between frames n and (n+m+1), and the entire procedure is repeated. A step-by-step description of the double-threshold scheme is given in Fig. 4.8.

For our earlier FAP scheme, the motivation to use the double-threshold scheme has been to improve the accuracy of parameter interpolation performed at the receiver between adjacent transmitted values. The same motivation does not hold for the above FIT method, since it explicitly considers interpolation error as part of its transmission strategy. Why, then, should one consider the FIT double-threshold scheme? The answer may be given as follows. Considering quantized parameter data, the FIT single-threshold scheme allows only integer thresholds. In effect, the double-threshold scheme may be viewed as equivalent to a single-threshold scheme that can allow a noninteger threshold. For example, the (0,1) double-threshold scheme produces average frame rate and speech quality that lie between those of the two single-threshold schemes with thresholds 0 and 1. This point will be more clear from the experimental results provided below.

Experimental Results

Below, we report experimental results obtained using the FIT method on the quantized pitch and gain data. Our speech data base consisted of a total of 11 utterances, representing about 25

(1) Transmit value at frame n

m ← 2

(2) Flag ← 0

k ← 1

(3) P ← (m-k)/m I_n + k/m I_{n+m}

E ← |[P] - I_{n+k}|

If E ≤ T2, go to (4)

n ← n+m-1

Go to (1)

(4) If E ≤ T1, go to (5)

Flag ← 1

(5) k ← k+1

If k ≤ m-1, go to (3)

(6) If Flag = 0, go to (7)

n ← n+m

Go to (1)

(7) (No transmission)

m ← m+1

Go to (2)

Fig. 4.8 Description of our FIT double-threshold scheme

seconds of speech, from 5 male and 5 female speakers. This data base is the same as the one used for computing average transmission frame rate data for our earlier FAP-type VFR schemes in [8].

Pitch: The FIT single-threshold scheme produced average frame rates of 35, 18 and 14 fps for values of the threshold T=0, 1 and 2, respectively. Using the (0,1) double-threshold scheme, we obtained an average frame rate of 26 fps. This latter rate should be compared against the rate of 35 fps that we had reported for our earlier (0,1) FAP scheme [8].

Gain: The FIT single-threshold scheme produced average frame rates of 57, 31 and 22 fps for values of the threshold T=0, 1 and 2, respectively. Using the FIT double-threshold scheme, we obtained average frame rates of 41, 26 and 19 fps for the two thresholds (T1,T2)=(0,1),(1,2) and (2,3), respectively. In contrast, the (2,3) double-threshold FAP scheme produced an average frame rate of 32 fps [8].

With the objective of not tolerating any speech quality loss, we have chosen to employ the single-threshold FIT scheme with the threshold T=0 for pitch transmission, and the (0,1) double-threshold FIT scheme for gain transmission. The use of the (0,1) double-threshold scheme for pitch and the (1,2) double-threshold scheme for gain yielded only a small speech

quality loss, which consisted mainly of occasional "roughness" (gain-related) and a couple of "clicks" (pitch-related) over the data base of 36 sentences given in Section 7.2.1.

4.4 Discussion and Recommendations

4.4.1 Transmission of Timing Information

With VFR transmission of a parameter, it is necessary to transmit timing information to indicate to the receiver the length of transmission interval between successive transmissions. To this end, we proposed in NSC Note No. 82 [7] that a 3-bit header be transmitted for every analysis frame. The first header bit is 1, only if pitch is transmitted for that frame; similarly, the second and third header bits are used to indicate if gain and LARs, respectively, are transmitted for that frame. This proposal of transmitting a 3-bit header allows the use of a separate transmission criterion for each of the three parameter groups: pitch, gain, and LARs, and hence accommodates one of the postulates of our perceptual model of speech.

4.4.2 Recommendations

Experimentally we found that the following VFR system yielded the maximum data compression without compromising speech quality relative to the full 100 fps fixed-rate system. In this experiment, an 11-th order LPC analysis was performed; the 11

LARs were quantized using 46 bits, which were allocated from the first to the 11-th coefficient as 6,5,5,4,4,4,4,4,4,3,3 bits; pitch and gain were quantized using 6 and 5 bits respectively; the simplified VFR scheme given in Section 4.2.8 was used for LAR transmission.

Recommended VFR System

LARs: Threshold, θ=1.0

Pitch: Single-threshold FIT scheme with threshold T=9 (see Section 4.3.2)

Gain: Double-threshold FIT scheme with thresholds T1=0 and T2=1 (see Section 4.3.2)

The above vocoder is referred to as PMH in Section 7.3.3, and it yielded the following transmission statistics for the 36-sentence (6 sentences x 6 speakers) data base given in Section 7.2.1. The LAR transmission frame rate computed over individual sentences varied from a maximum of 44 fps to a minimum of 14 fps, with an average of 31 fps. The average, maximum and minimum transmission frame rates for pitch were: 34, 43 and 25 fps respectively, and those for gain were: 40, 54 and 24 fps respectively. The bit rate varied from a maximum of 2817 bps to a minimum of 1274 bps, with an average of 2120 bps. This average bit rate of 2120 bps for the above VFR system should be contrasted with the bit rate of 5700 bps for the full 100 fps fixed-rate system. With the

benefits of Huffman coding and variable order linear prediction [1], the average bit rate would be further reduced to about 1600 bps for continuous speech. With explicit silence detection, an average bit rate of less than 1000 bps may be achieved for normal conversational speech.

A formal subjective speech quality test was conducted to evaluate the effectiveness of our perceptual-model-based VFR system. The results of this study are given in Section 7.3.3. Specifically, we found that the above recommmended VFR system produced speech quality which equalled or surpassed the full 100 fps fixed-rate vocoder.

5. SYNTHESIS

In this section, we report the results of our work on the following three items: optimal linear interpolation of synthesizer parameters, gain implementation, and all-pass excitation.

5.1 Optimal Linear Interpolation

In narrowband LPC speech compression systems, the process of parameter interpolation at the receiver helps in smoothing the roughness in the synthesized speech which is normally associated with infrequent parameter updating. Simple linear interpolation (SLI) has been used almost exclusively in these systems. In an earlier study we found that the spectral error due to interpolation was much larger than the error due to quantisation [1]. This result suggests that better parameter interpolation approaches than the simple linear scheme should be investigated. With this motivation, we developed an optimal linear interpolation (OLI) scheme that requires the transmission of an extra parameter per data frame, $:0 \leq \alpha \leq 1$. The value of α is determined as that point along the line used for linear interpolation which is closest (in the mean square sense) to the point determined by the actual parameter values at the instance where interpolation is desired. The transmission of α requires 50-150 bits/sec, depending on the frame rate and the number of bits used for quantizing α.

Theoretical and experimental results that we obtained with the new interpolation scheme have been presented in detail in a BBN report [11], which was also issued as NSC Note No. 59. Briefly, theoretical results showed that in the space of parameter vectors, the OLI scheme corresponds to an orthogonal projection of the actual parameter vector at the interpolation point onto the line passing through the two parameter vectors that are used in the interpolation. Several ways of using the OLI scheme with a variable frame rate transmission system are also given. Experimental results showed that the OLI scheme improved speech quality relative to the SLI scheme, especially during rapid transitions in the speech signal. In addition to informal listening tests, we investigated the waveforms and spectrograms of synthesized speech with OLI, and the time history of the spectral error. In our experience, the optimal scheme is most advantageous when used with low bit rate, variable frame rate transmission systems.

5.2 Gain Implementation

We investigated three issues involving linear predictor gain parameter. The first issue was the choice of the gain parameter for transmission; we discussed this issue in Section 3.3. The second issue considered the problems associated with implementing the speech signal energy as a multiplier at the output of the synthesizer filter instead of the more commonly used method of

applying it at the filter input. The third issue was the treatment of cases for which speech signal energy had values less than 1 (or negative when considered in decibels). Below, we describe our work on the second and the third issues.

The use of the normalized filter [4] (see Section 3.3) is recommended for implementation of the synthesizer on the SPS-41 for many reasons, such as better round off noise and scaling properties, the availability of sine and cosine tables in the SPS-41, etc. Placing the gain multiplier at the output of the normalized filter rather than at its input serves to alleviate dynamic range problems. However, care has to be exercised in implementing the speech signal energy at the output of either the normalized filter or the regular filter. The difficulty, implied in the above statement, arises from the nonzero initial conditions of the filter. Whenever there is a relatively large change in speech signal energy from one frame to the next, say, of the order of 10 dB, then the synthesized speech is found to have signal amplitudes quite different from those of the original input speech. For example, in an unvoiced-voiced transition, the first voiced frame in the synthesized speech has relatively large signal amplitudes compared to the original speech. We showed both experimentally and mathematically that this problem was due to the nonzero initial conditions of the filter. When listening to speech synthesized with speech signal energy implemented at

the output of the synthesizer filter, we perceived these distortions in signal amplitudes as annoying "knock sounds". A solution to the problem, which we found to be satisfactory, is to zero the initial conditions whenever the absolute frame-to-frame energy change exceeds a given threshold (about 12 dB). With this method, the distortions in signal amplitudes which caused the perception of "knock sounds" were eliminated.

In logarithmically quantizing speech signal energy we used a range of Ø to 45 dB. Any signal energy less than Ø dB was quantized as Ø dB. From synthesis experiments we found that this strategy of raising the energy from a negative dB value to Ø dB produced relatively large perceivable noise during stop sounds, pauses and silences. This led us to quantize energy values less than or equal to Ø dB as a given negative dB. We found through listening tests that when we used a large negative dB value, the beginnings of certain speech sounds (e.g., [h], [n], [d]) were somewhat cut off. By experimentation, we found a value of -3 or -4 dB to be satisfactory.

5.3 All-Pass Excitation

With the use of the pulse/noise excitation for the minimum-phase LPC synthesizer, the synthesized speech was found to have larger peak amplitudes than the natural speech used in the analysis. To accommodate this situation, we have used 9 bits

to store input or natural speech samples, and 12 bits to store synthesized speech samples. Since the full dynamic range possible with 12 bits was not effectively used in storing the synthesized speech samples, the signal-to-noise (noise at the D/A converter) ratio was lower, producing sometimes less desirable audio quality at the output of the D/A converter. To overcome this problem, we employed an all-pass excitation as described below.

We chose an 8th order all-pass filter given in [12], which was specifically designed to minimize the peak amplitude of its impulse response. All-pass excitation signal can be obtained by filtering the pulse/noise excitation signal through this all-pass filter. To simplify computations, however, we precomputed once at the start 40 samples (4 ms at 10 kHz sampling rate) of the impulse response of the all-pass filter and stored them in memory. If a given frame was unvoiced, we used the random noise sequence directly as the excitation signal (i.e., no all-pass filtering was done); this strategy worked fine since high peak amplitudes occurred only in voiced speech. For a voiced frame, we chose one of the following two cases, depending on the value of the pitch period for that frame: 1) If pitch period was longer than 4 ms, we took the 40 samples of the all-pass impulse response and appended at the end with the required number of zeros to generate the excitation signal. 2) If pitch period was

shorter than 4 ms, we used the "aliased" version of the 40-sample sequence which was obtained by considering the periodic occurrence of this sequence at a rate given by pitch frequency.

By conducting synthesis experiments, we found that peak amplitudes were in fact lowered when using the specific all-pass excitation discussed above. Even in this case, however, peak amplitudes of synthesized speech were higher than those of the natural speech; the increase in peak amplitudes due to synthesis was often found to be about 6 dB or less. We accommodated this increase by using 11-bit natural speech samples, and 12-bit synthesized speech samples. Using this approach, the audio quality of speech at the output of the D/A converter was found to be better than what we had previously found. We used this approach in generating stimuli for subsequent subjective quality tests.

6. A MIXED-SOURCE MODEL

We developed a new model for generating the excitation signal for the synthesizer of the narrowband LPC vocoders, with the objective of enhancing the naturalness of the synthesized speech. Most present-day narrowband vocoders employ an idealized source (or excitation) model, which is either a sequence of quasi-periodic pulses for voiced sounds, or white noise for unvoiced sounds. This voiced/unvoiced model seems to be largely responsible for the "buzziness" and lack of naturalness perceived in the resulting synthesized speech. Our new source model, called mixed-source model, combines both pulse and noise sources in a novel way. Based on the observation that spectra of voiced speech sounds (e.g., voiced fricatives and even certain vowels) exhibit devoiced or incoherent high frequency bands, the model divides the spectrum into a low frequency region and a high frequency region, with the pulse source exciting the low region and the noise source exciting the high region. The cutoff frequency F_c that separates the two regions is adaptively varied in accordance with the changing speech signal.

The mixed-source model is described in detail in a paper which is included in this report as Appendix 7. As depicted in Fig. 4 of that paper, the outputs of the low-pass and high-pass filters are added, multiplied by the source gain and applied to the synthesizer as the excitation signal. For unvoiced sounds

(F_c=0), the model employs a pure noise excitation. Since small changes in F_c are not perceptible, it is sufficient to quantize F_c into 2-3 bits for transmission purposes.

The cutoff frequency F_c is a continuous parameter, and so errors in the extraction of F_c degrade the quality of the synthesized speech much more gracefully than the errors in the binary voicing parameter of the voiced/unvoiced model. Thus, the mixed-source model promises to be a more robust source model.

We developed a method for automatically extracting F_c from the speech signal. The method is a peak-picking algorithm on the signal spectrum. It determines periodic regions of the spectrum by examining the separation between consecutive peaks and determining whether the separations are the same, within some tolerance level. F_c is taken to be the highest frequency at which the spectrum is considered periodic.

In our implementation of the mixed-source model, we rounded the automatically extracted value of F_c to the nearest 500 Hz. Therefore, we needed low-pass and high-pass filters with cutoff frequencies separated by 500 Hz. For each value of F_c, the 3 dB points for the low-pass and high-pass filters were designed to be equal to F_c, in order that the spectrum of the final excitation may be as flat as possible. The roll-off of the filters was considered to be of secondary importance, but should not be very

sharp in any case. We considered FIR (finite impulse response) as well as recursive (low order Butterworth) filter designs. The filter designs were stored and used in the synthesis as the need arose. Both FIR and recursive filter designs gave similar perceptual results.

Results of synthesis experiments conducted to test the effectiveness of the mixed-source model are given in Appendix 7. Briefly, the model was found to largely eliminate the "buzzy" quality of vocoded speech, perform better for female speech, and result in a certain "fullness" in perceived speech quality that was absent with the voiced/unvoiced synthesis.

7. SUBJECTIVE SPEECH QUALITY EVALUATION

7.1 Introduction

We describe our work on subjective quality evaluation in three parts. Section 7.2 describes the development and testing of an improved method for measuring subjective quality, using our Phoneme-Specific sentence materials. Section 7.3 describes three applications of this method to practical problems: (1) Determining parametrically how subjective quality depends on vocoder parameters, specifically a) the order of the linear predictor (number of poles), b) the step size for quantization of the LPC coefficients (log area ratios or LARs), and c) frame transmission rate. In addition to their usefulness in vocoder design decisions, these data were also needed for the development of our objective method for assessing speech quality (see Section 8). (2) Proving that a given reduction of bit rate is achieved at a much smaller cost in reduced quality if the bit rate is reduced by substituting a variable for a fixed transmission schedule, rather than reducing the predictor order, or coarsening quantization of the coefficients. (3) Demonstrating formally the superior quality and low bit rate of our perceptual-model-based VFR scheme described in Section 4.2.

Finally, Section 7.4 describes miscellaneous topics such as: (1) a phoneme-specific intelligibility test, using nonsense

materials, which we later decided was not appropriate except for testing LPC systems which had been implemented in real time, which ours had not; (2) the effect of lost packets on the intelligibility of speech transmitted over ARPANET; (3) development of an inventory of descriptors for different perceptual attributes of LPC vocoder speech quality; and (4) an attempt to reduce the effects of stimulus sequence on listeners' judgments.

7.2 Development of Method

7.2.1 Phoneme Specific Sentences

The development of our Phoneme-Specific test sentences grew from the observation that different vocoders may cause different types of quality degradations within a single test sentence. For example, one system may degrade the nasal consonants, and another the fricatives. Such differences are a major cause of the variability commonly found in subjective quality testing. If such information could be made explicit, it would have important diagnostic implications for how a vocoder should be modified to improve its quality.

Judgments of global quality are not easy when the stimuli being compared differ in a variety of ways. Nor is it easy to compare speech samples with respect to some particular property

when they differ with respect to many other properties as well. Further, the psychometric literature is unequivocal in showing that subjects find quantity much easier to judge than quality. One way to simplify the subject's task is to arrange that the stimuli presented for judgment differ with respect to only one perceptual dimension at a time. Note that this is not the same as asking the subject to attend to only one perceptual dimension at a time, when they differ in other ways as well.

We attempted to achieve this perceptual effect, or something close to it, by analyzing the sources of distortion introduced into speech by the LPC vocoding process, and targetting each of these sources with one or more sentences designed to maximize the errors due to it, while simultaneously minimizing the errors due to the other sources. Although our tests were aimed specifically at LPC vocoders, they are probably applicable to other methods of vocoding as well. The resulting sentences are Phoneme-Specific, in that they concentrate all phonemes with similar acoustic properties in a single sentence. This contrasts with earlier materials, which treated any sentence as equivalent to any other. The equivalent-sentence paradigm involves a logical inconsistency because it implicitly assumes that speech is homogeneous, and at the same time denies this assumption in its attempt to achieve phonetic balance by forcing the relative frequency of occurrence of phonemes within the test materials to match those of the language at large.

There are three primary sources of distortion inherent in linear predictive vocoders. The first derives from the predictor model itself. The linear predictor coefficients effectively define the spectrum of an all-pole filter, which is adjusted until it best matches the envelope of the power spectrum of a short sample of input speech. Some speech sounds, however, are not adequately modelled by an all-pole spectrum, since their spectra contain zeroes as well as poles (although adequate matches can be obtained if the number of poles is sufficiently large). Errors deriving from this source degrade phonemes whose spectra contain zeroes, such as nasals and fricatives. The second source of distortion is in the quantization of the LPC coefficients for transmission. The quantization introduces some inaccuracy to the degree that the spectrum specified by the quantized coefficients differs from that specified by the same coefficients before quantization. Distortions due to this source should be most apparent when the speech spectrum is changing relatively slowly, as in vowels and semi-vowels. Third, the time interval defining the waveform sample is moved down the waveform by a time equal to the reciprocal of the frame rate, and the spectral modelling is repeated. The slower the frame rate, the wider the intervals at which the speech spectrum is sampled, and the greater the chance that rapidly changing parts of the waveform will be inadequately represented. This type of error should be most noticeable when a system with too slow a frame

rate has to process speech containing stops and affricates, which are characterized by rapid changes in both spectrum and amplitude.

A set of four phoneme-specific sentences was selected from a much larger set, which appears complete in Appendix 8. The four phoneme-specific sentences were intended to target the sources of error described above. Two additional "general" sentences were included, which contained many consonant clusters and unstressed syllables. The six sentences are as follows:

1. Why were you away a year, Roy?
2. Nanny may know my meaning.
3. His vicious father has seizures.
4. Which tea-party did Baker go to?
5. The little blankets lay around on the floor.
6. The trouble with swimming is that you can drown.

The first four sentences include among them all the consonants of English, except /l/, /θ/, and /j/. The first sentence, contains only vowels and glides. These sounds have all-pole spectra, which change slowly, and contain no abrupt changes in level. This results from the fact that these sounds are produced with a relatively open vocal tract, excited at the bottom, and without any shunting cavities to cause zeroes (as in /l/) or extra formants. Furthermore, all the sounds are voiced, and only slow changes of pitch occur.

The second sentence contains only (nasalized) vowels and nasals. It is therefore also voiced throughout, and its spectrum and level change relatively slowly. Both the nasals and the nasalized vowels contain zeroes in their spectra, however, which should create problems for LPC vocoders in the spectral matching stage.

Besides vowels, the third sentence contains only voiced and unvoiced fricatives. Fricatives contain zeroes in their spectra (actually pole-zero pairs which approximately cancel each other), but have spectra very different from those of voiced sounds, due to the noise excitation. Rates of amplitude change are still slow, since affricates were excluded from the sentence.

The fourth sentence contains only vowels and all the stops and affricates except /j/. The spectrum and amplitude of the speech wave change frequently and abruptly, and there are many voiced/unvoiced transitions. This sentence should maximally strain a vocoder's ability to follow rapid changes.

The last two sentences were included as "general, non-diagnostic" sentences, partly to include problematical clusters which would have sullied the purity of the phoneme-specific sentences, but also to increase the number of rapid unstressed (and reduced) syllables, which tend to be less clearly articulated.

A second set of dimensions along which samples of speech can vary concerns the idiosyncratic differences among speakers. Following arguments similar to those above for sentence materials, it is important to represent a wide a range of speaker's physical (as opposed to dialectal) characteristics rather than to choose "typical" speakers. Therefore, we recorded twenty talkers, ten male and ten female, reading each of the sentences, and selected from these three males and three females so as to retain the full range of fundamental frequency and nasality found in the whole group. (Nasality was measured by an accelerometer mounted on the talker's nose, whose output was compared in the second, nasal sentence, and the fourth, non-nasal sentence, with overall levels equated.) Talkers who spoke slowly, or had regional accents, were eliminated.

7.2.2 Psychophysical Method

A variety of different psychophysical tasks can be used for assessing subjective speech quality. These represent different compromises between the complexity and duration of the subject's task. The subjective task that imposes fewest constraints on the listener is the paired comparison task. Pairs of stimuli are presented, and the listener simply indicates which member of each pair he prefers. Alternatively, he may assign numbers to show how similar the two stimuli appear, yielding similarities or proximity data. Since only two stimuli are presented at a time,

the listener never has to explicitly resolve the problem that the members of successive pairs may differ in different ways. Unfortunately, the number of paired comparisons that has to be made increases with the square of the number of stimuli, so that the exhaustive procedure becomes unmanageable when there are more than 15 to 20 stimuli to be compared, although various sampling schemes are available. Thus, paired comparisons are easy but tedious.

An alternative is a ranking task, in which subjects are given several stimulus sentences, and have to rank order them according to quality. With conventional materials this is a very difficult task that generates much variability, since the subject must decide how to trade off one sort of degradation against another, and apply that trade-off consistently. When the speech varies along several perceptual dimensions, maintaining consistency with respect to each of the required trade-offs becomes impossible. The foregoing difficulties can be reduced by using the Phoneme-Specific sentence materials described above, and presenting stimuli for ranking that consist of only a single sentence processed by all the vocoder systems to be compared. As compared with the paired-comparison task, this reduces the amount of data to be collected, at the expense of making the listeners task slightly more difficult, and introducing the risk of reduced reliability. At the same time, the ranking task retains some of

the desirable features of the paired comparison task. Since a stimulus can be listened to repeatedly, the subject can build up his rank order by starting with a pair, then placing the third stimulus in the correct ranking with respect to the preceding two, and so on. Thus, new stimuli may be added by a series of paired comparisons with the members of the existing rank order. The rank-order procedure reduces the number of times each stimulus must be presented to perhaps five or ten per stimulus.

A complication of the ranking task results from the fact that the range of qualities encountered within a single test sentence as processed by several vocoders may be very different from the range for a second test sentence. Thus there may be a considerable range of qualities associated with the lowest rank, but there is no way for the subject to express this, although he might be willing and able to do so if given the chance. The ranking task becomes more difficult, and the data from it become less reliable, as the number of systems to be ranked increases. The method is probably not appropriate when more than 20 systems are to be compared.

A rating task avoids some of these problems, and is perhaps the most efficient task possible, in that it requires only a single presentation of each stimulus. In practice, several presentations are often used, to improve reliability. At the same time this task requires most from the subject, who must

assign numbers that reflect his perception of quality, and stick to the same rating system through the whole experiment. His criterion may drift during an experiment lasting an hour or more, and it is difficult to assess how much drift has occurred, and to correct for it.

An important question is whether the tasks outlined above (and other possible tasks too) force subjects to perceive the stimuli differently, or whether each subject makes use of a single underlying perceptual structure to perform all tasks. If the latter could be demonstrated, it would allow the task to be selected on the basis of convenience for any given application.

7.2.3 Multidimensional Scaling and Analysis

It should be clear from the arguments above that speech quality can vary along several perceptual dimensions simultaneously, and that these may be separable, especially if phoneme-specific sentences are used as test materials. Furthermore, diagnostic information about different aspects of quality can be derived from such data. Yet most approaches to quality assessment start with the assumption that quality is a unidimensional variable, thus ignoring the diagnostic potential. Unidimensional testing also introduces a major source of inter-subject variability, since it requires the subject to collapse his multidimensional percepts onto a single dimension to

arrive at a response, and different subjects may weight the various perceptual dimensions quite differently.

A major justification for treating quality as a unidimensional variable is that it will be used for choosing the best of a set of candidate systems, an inherently one-dimensional task. But counterexamples can easily be found: a vocoder that yields excellent quality for female voices, but fails disastrously on male voices will not receive a high quality rating on a unidimensional scale. Yet it may be ideally suited for an application in which only females will use it, a fact a unidimensional test would not discover. It would seem to be better to recognize that quality is multidimensional, and collect and analyze data accordingly, and only later collapse the multidimensional result onto a unidimensional scale if desired. Among other things, this would permit the tester to decide how to combine the various perceptual dimensions, rather than be forced to accept the idiosyncratic combinations adopted by the subjects in a unidimensional task.

Multidimensional scaling (MDS) methods attempt to model empirical data by representing each stimulus, or vocoder system, as a point in an n-dimensional space, such that the data reconstructed from the model match the empirical data as closely as possible. There are several classes of models, which are hierarchically related in that each class is a special case of

the next-higher class in the hierarchy. The simplest is the vector model. The stimuli (here the different vocoders) are represented by points in an n-dimensional space, and each condition under which data are collected (different sentences or subjects) can be represented by a vector through the space. The data are represented by the ordering and relative spacing of the stimulus points as projected onto the appropriate vector . An example of a vector model appropiate to scaling preference data is MDPREF [16].

A second type of model appropriate to speech quality assessment is the weighted Euclidean model, typified by INDSCAL [17]. INDSCAL was developed to model explicitly the large individual differences in how subjects perceive stimuli. The model assumes that all subjects use the same set of underlying perceptual dimensions, but that the relative salience of these dimensions varies among subjects. Therefore, INDSCAL models each stimulus as a point in a "group space" of one or more dimensions, which represent the perceptual dimensions common to all subjects. To model the data for a particular subject, the dimensions of the group space are linearly stretched or shrunk until they reflect the relative salience of the dimensions for that subject. Thus the INDSCAL solution consists of sets of coordinates for the stimuli in the group space, and sets of weights for deforming the group space to produce the idiosyncratic space for each subject.

The distance between stimulus points in the space represents their similarity: stimuli that are judged very similar are represented by points that are very close together in the space.

The multidimensional space that is used to model the data in these examples is a perceptual, or subjective space. The analysis itself does not identify the factors represented by the coordinate axes of the space, which are simply those that give the best match to the input data. Often, but not always, the axes can be identified from the way the stimuli are distributed in the space. Otherwise, several additional psychophysical experiments may be required to identify the axes. Even this does not guarantee that the axes will be identified: sometimes no objective properties of the stimuli can be found that correspond to particular subjective dimensions. Unfortunately, these shortcomings reduce the usefulness of multidimensional scaling for routine quality testing, although they can be highly beneficial in development work.

Preference data, such as is generated by the rating or ranking tasks described above, can be analyzed by vector models such as MDPREF, or by weighted Euclidean models such as INDSCAL if the data are first converted into proximities [17].

7.2.4 A Test of the Method Using 2600 bps Systems

To try out our method, we selected a set of 12 LPC vocoder systems, whose bit rates were equated to 2600 bps to test the method's ability to discriminate small differences of quality. Each of the 36 test sentences (6 phoneme-specific sentences x 6 talkers) was low-pass filtered at 5 kHz and processed through each system. Each system used 9, 11, or 13 poles; and inter-frame intervals (reciprocal of frame rate) were 25, 20, or 15 ms, or variable based on data analyzed every 10 ms. Details of the parameter combinations for each system appear under Fig. 7.1. In the five variable rate systems, frames of spectral data were analyzed every 10 ms, but each frame was transmitted only if the spectral difference between it and the previous transmitted frame exceeded a threshold. The quantization step size of the LARs, and, for the VFR systems, the threshold, were adjusted so that the overall bit rate of all systems was equated at 2600 bps, averaged over the 36 test sentences. Quantization step size varied between 0.2 dB and 1.75 dB, and the VFR thresholds varied between 1.0 and 1.75 dB, yielding average frame rates between 47 and 31 per second. Pitch and gain were coded in 6 and 5 bits respectively, and were transmitted at the same frame rate as the coefficients for the fixed-rate systems, but at a constant rate of 50 fps for the VFR systems, to avoid confounding excitation and spectral variables. One final vocoder used 13

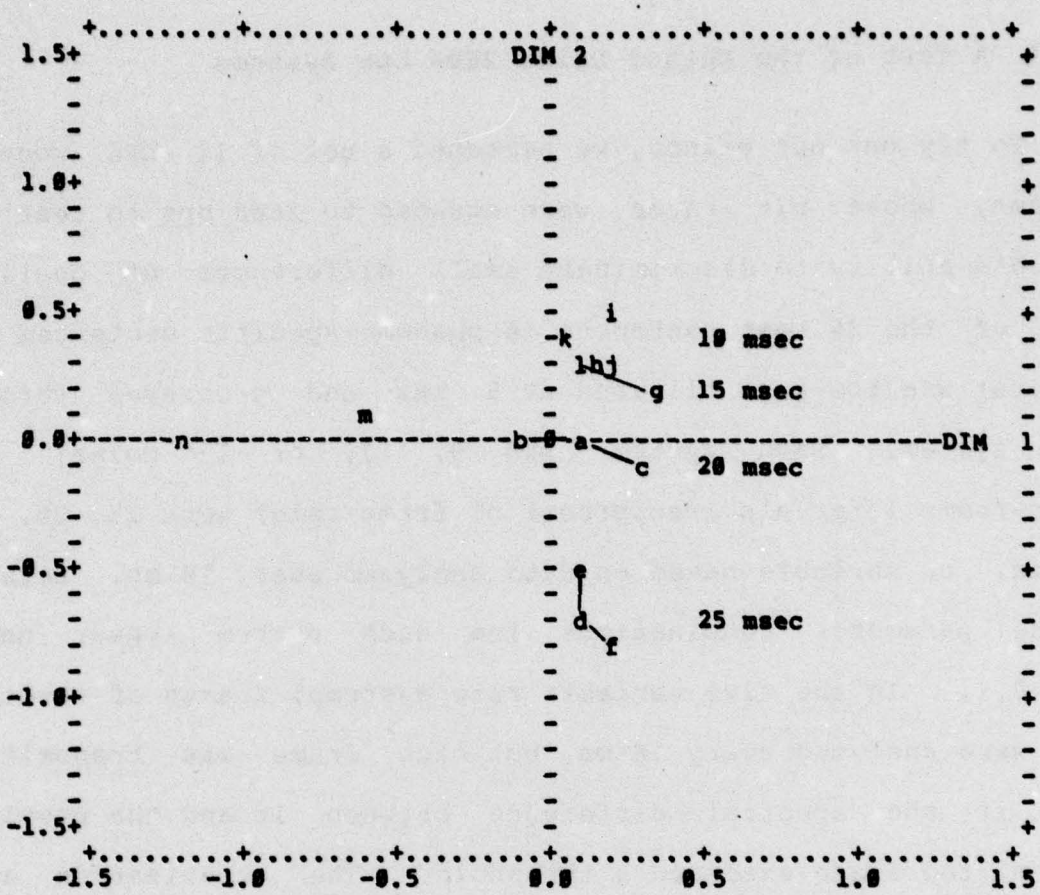


Figure 7.1: Projection on 1-2 plane of points representing Vocoder systems in 3-dimensional MDPREF solution.

SYSTEM	# POLE	FRAME SIZE	VAR-RATE THRESH	FRAME RATE	STEP SIZE	BITS PER SECOND EXPECT	BITS PER SECOND FOUND
a	11	20	-	50	1.0 dB	2650	2630
b	9	20	-	50	0.6	2650	2633
c	13	20	-	50	1.4	2700	2681
d	11	25	-	40	0.45	2640	2610
e	13	25	-	40	0.7	2640	2612
f	9	25	-	40	0.2	2680	2652
g	9	15	-	67	1.75	2666	2618
h	11	10	1.5 dB	35	0.5	2660	2574
i	11	10	1.0	46	1.0	2650	2652
j	11	10	1.75	32	0.25	2627	2687
k	13	10	1.5	38	0.6	2685	2766
l	11	15	1.5	33	0.4	2600	2535
m	13	10	-	100	0.0	(UNQUANTIZED)	
n	ORIGINAL WAVEFORM, DIGITIZED AND RECONSTITUTED (ie PCM)						

poles, with unquantized coefficients, and an inter-frame interval of 10 ms. Finally, the digitized but unprocessed waveforms were included to act as undegraded anchors. The unprocessed speech was effectively 110 kbps PCM.

The same four subjects served in two judgment tasks, one a ranking task and the other a rating task. Our purpose in collecting data with two different psychophysical methods was to test the idea that any judgments required of a subject are made on the basis of a single underlying perceptual structure, or set of psychological dimensions. If both tasks give similar results, this idea is supported, and the most efficient task may then be selected for subsequent experiments.

In total, there were 504 stimulus sentences--36 test sentences x 14 systems (12 vocoders with quantized coefficients, 1 with unquantized coefficients, and 1 PCM). For the rank-ordering task, these were transferred to Bell and Howell Language Master cards, to permit random access. Each subject rank ordered the 14 versions of a given test sentence, separately for each of the 36 sentence-speaker combinations, which were arranged in a different counterbalanced order for each subject. The task was self-paced, and took a total of 6 to 9 hours, spread over several days.

For the rating task, the 504 stimulus sentences were recorded into a carefully counterbalanced order, in which each sentence, speaker, and system followed every other sentence, speaker, and system (except itself) as nearly the same number of times as possible. Stimuli were presented in blocks of 10; the first stimulus in each block repeated the last stimulus in the preceding block, and was not scored. Also, unbeknownst to the subjects, the first 10 blocks of stimuli were repeated at the end, thus permitting an assessment of consistency and drift. Consistency was high and drift was negligible. The four subjects had also served in the ranking task; they assigned "degradation ratings" to each stimulus, with higher numbers representing more degradation (lower quality). Subjects were told to assign zero degradation to any undegraded sentences they heard, and to try to assign ratings on a proportional basis, with twice as large a number representing twice the degradation, as in a magnitude estimation task with a natural zero. Since the first few judgments from each subject effectively determined his step size, each subject's ratings were later normalized by dividing through by his mean rating. The rating task took just over an hour.

Results

The data from each task were pooled across subjects, and analyzed separately with MDPREF [16]. The first three dimensions accounted for 70.4%, 8.9%, and 6.0% respectively of the variance

in the rating data, and 65.8%, 11.9%, and 7.6% of the variance in the rank data. In each case, the fourth dimension accounted for only an additional 3% of the variance. Canonical correlation [18] of the two solutions showed them to be almost identical, with the first three (orthogonal) linear composites correlating at 0.988, 0.930, and 0.758 respectively. The first two of these are significant well beyond P<.001 and the third is significant at P<0.01 (chi-square = 69.6, with 9 df; 30.0, with 4 df; and 8.98, with 1 df). The conclusion that rating and ranking tasks produced virtually identical results seems justified, which means that the more efficient task (rating) can be used in future assessments.

Figures 7.1 and 7.2 show the distribution of the vocoder systems in the 3-dimensional solution space, as projected onto the Dim 1 x Dim 2 plane, and the Dim 1 x Dim 3 plane respectively. Each test sentence on which ratings were obtained would be represented by a vector through the space, but they are not shown, to avoid cluttering the figure (more detail can be found in [19]). The relative performance of two vocoders on a particular speaker-sentence combination is represented by the relative positions of the projections of the points representing the systems onto the corresponding vector.

The results show a clear separation of the systems as a function of 1) the number of poles, and 2) the inter-frame

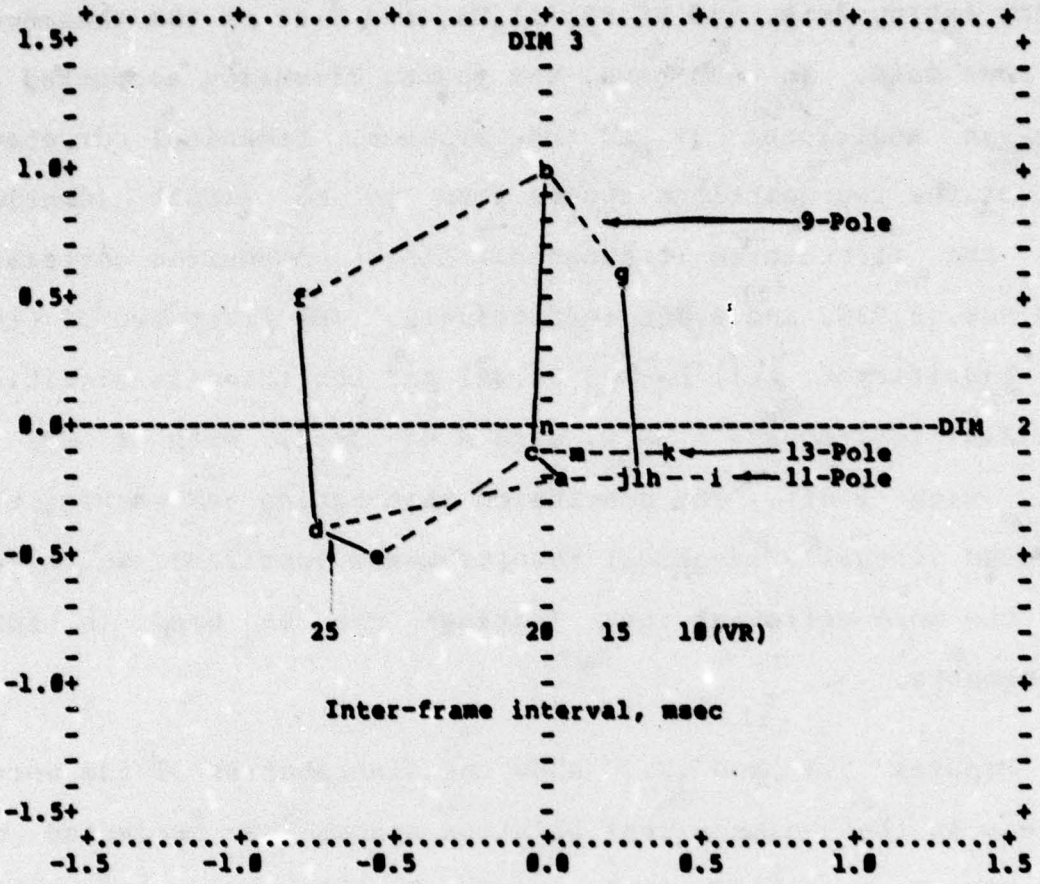


Figure 7.2: Projection on 2-3 plane of points representing Vocoder systems in 3-dimensional MDPREF solution.

SYSTEM	# POLE	FRAME SIZE	VAR-RATE THRESH	FRAME RATE	STEP SIZE	BITS PER SECOND EXPECT	FOUND
a	11	20	-	50	1.0 dB	2650	2630
b	9	20	-	50	0.6	2650	2633
c	13	20	-	50	1.4	2700	2681
d	11	25	-	40	0.45	2640	2610
e	13	25	-	40	0.7	2640	2612
f	9	25	-	40	0.2	2680	2652
g	9	15	-	67	1.75	2666	2618
h	11	10	1.5 dB	35	0.5	2660	2574
i	11	10	1.0	46	1.0	2650	2652
j	11	10	1.75	32	0.25	2627	2687
k	13	10	1.5	38	0.6	2685	2766
l	11	15	1.5	33	0.4	2600	2535
m	13	10	-	100	0.0	(UNQUANTIZED)	
n	ORIGINAL WAVEFORM, DIGITIZED AND RECONSTITUTED (ie PCM)						

interval, of the vocoders. Furthermore, the separation along these two dimensions was orthogonal, suggesting that the perceptual effect of changing the number of poles ("static" spectral accuracy) was independent of the perceptual effect of changing the inter-frame interval ("dynamic" spectral accuracy). The orientation of the test-sentence vectors in the space showed that the separation of the fixed-rate systems by inter-frame interval was achieved as a result of the specially composed sentence materials, with the short inter-frame interval systems performing better on the rapidly changing sentence (No. 4: Which tea party...), and the long inter-frame interval systems, with more bits per frame, doing better on the slowly-varying sentences (Nos. 1 and 2). The VFR systems were located correctly for their inter-frame intervals of 10 ms, but performed unexpectedly badly on the slow-moving sentences, Nos. 1 and 2. Separation of the vocoders as a function of the number of poles resulted from the use of the different talkers, with the relative performance of systems with 13, 11, and 9 poles on a particular sentence being highly correlated with the mean fundamental frequency of the talker in that sentence. Nine-pole systems performed almost as well as 11- or 13-pole systems on high-pitched talkers, but not on low-pitched talkers.

Conclusions

1) Phoneme-specific sentences, when used in a rating task to assess the subjective quality of a set of twelve very similar LPC vocoders, were able to distinguish quite small differences in quality. The data were both reliable and diagnostically useful, in that they permitted the particular parameter causing degradation to be identified.

2) Virtually identical MDPREF solutions were obtained for rating and rank-ordering tasks, which strongly supports the idea that subjects used the same set of perceptual dimensions when responding to vocoder-processed speech samples, for both of these tasks. This means that the most cost-effective task can be used exclusively -- in this case the rating task.

7.3 Applications of the Method

7.3.1 Effects of Vocoder Parameters on Quality

A factorial subjective-quality study was performed to measure how the quality of LPC vocoded speech is affected by three different methods of reducing bit rate. A paper on this study was presented at the 1977 ICASSP Conference at Hartford, Conn., and is reproduced as Appendix 9. The three methods of reducing bit rate were:

1) reducing the number of poles (P) used for spectral matching,
2) coarsening quantization step size (Q) for the LAR coefficients,
3) reducing the frame transmission rate (R).

The set of spectral parameter values that were used are shown below, together with the number of bits per frame.

Quantization Step Size, Q	No. of Poles, P 13	11	9	8
0.25 dB	76	--	--	--
0.5 dB	63	55	47	43
1.0 dB	50	44	38	35
2.0 dB	37	33	29	27

Bits per frame, excluding pitch and gain, for all combinations of number of poles and quantization step size used in the present study.

Each combination of spectral parameters (except 13 poles with 0.25 dB quantization) was combined with four different fixed transmission rates of R = 100, 67, 50, and 33 fps, yielding 48 LPC systems (4x3x4). Two additional systems were included: an LPC system with 13 poles, quantization step size of 0.25 dB, and transmission rate of 100 fps; and PCM speech at 110 kbps (i.e. the 5 kHz bandwidth speech sampled at 10 kHz and quantized to 11 bits), to act as an undegraded anchor. Pitch and gain were coded in 6 and 5 bits respectively, and transmitted at the same frame rate as the coefficients. The measured overall bit rates of the LPC systems ranged from 8430 bps (P = 13, Q = 0.25 dB, R = 100 fps), down to 1225 bps (P = 8, Q = 2.0 dB, R = 33 fps), as shown

in Table 7.1. (These rates do not include the benefits of Huffman coding.)

Our earlier subjective quality tests showed the necessity of passing all sentence materials through all systems. Unfortunately, we could not use all 36 speaker-sentence combinations in the present study, since passing them through all 50 vocoder systems would have made the study unmanageably large. We therefore selected a subset of seven speaker-sentence combinations, and confirmed that they were adequately representative of the full set by showing that the MDPREF solution obtained from the data from the subset was substantially the same as that obtained from the complete set. (Canonical correlations between the first three linear composites of the two solutions were 0.991, 0.954, and 0.923.) The selected sentence tokens were: JB1, DD2, RS3, AR4, JB5, DK6, and RS6 (the initials identify the speaker and the number identifies the sentence). Average fundamental frequency is shown for each test sentence in the second row of Table 7.1.

Each of the seven input sentences was low-passed at 5 kHz, digitized (11 bits, 10 kHz), and passed through each of the 50 simulated vocoder systems, to yield a total of 350 different stimulus items. A counterbalanced presentation sequence was generated, in which each of the 50 systems followed every other system once, and each speaker and sentence followed each other

Sentence ID:					JB-1	DD-2	RS-3	AR-4	JB-5	DK-6	RS-6	
Sentence F0 in Hz:					119	134	195	165	124	97	193	
ID	#P	QdB	fps	kbps	Ratings:		by	sentence				pooled
000		(PCM)		110.0	14.4	16.3	17.1	6.4	7.9	11.5	11.9	12.22
111	13	.25	100	8.430	44.2	44.3	52.9	51.6	28.1	34.9	46.4	43.22
121	13	0.5	100	7.160	54.0	49.8	58.3	52.0	30.6	34.9	51.3	47.27
122	13	0.5	67	4.768	51.5	42.7	66.6	58.3	33.7	32.7	53.4	48.42
123	13	0.5	50	3.576	45.6	50.9	72.0	60.1	31.9	34.1	53.6	49.75
124	13	0.5	33	2.381	50.2	52.1	72.8	67.7	52.9	45.7	62.3	57.68
131	13	1.0	100	5.930	60.6	45.9	59.6	54.7	37.9	34.7	62.7	50.86
132	13	1.0	67	3.947	57.7	53.2	63.4	59.8	38.7	30.0	55.1	51.12
133	13	1.0	50	2.960	51.1	53.6	70.9	58.7	34.5	32.9	57.7	51.34
134	13	1.0	33	1.971	51.1	52.4	73.0	70.9	57.8	45.8	68.8	59.96
141	13	2.0	100	4.930	70.3	63.5	68.4	62.5	56.2	53.7	66.6	63.04
142	13	2.0	67	3.280	71.0	63.5	70.5	61.2	55.6	59.3	59.7	62.97
143	13	2.0	50	2.460	71.8	63.2	72.8	60.0	47.1	43.8	64.6	60.46
144	13	2.0	33	1.638	59.9	64.1	75.1	70.6	52.4	44.2	66.0	61.78
221	11	0.5	100	6.430	56.3	50.4	54.9	50.6	29.6	37.7	54.0	47.65
222	11	0.5	67	4.280	51.8	48.3	67.2	57.6	39.9	30.2	52.1	49.58
223	11	0.5	50	3.210	55.0	52.4	68.9	63.1	38.6	34.9	59.4	53.19
224	11	0.5	33	2.138	49.0	55.1	73.4	65.0	48.4	41.0	66.0	56.86
231	11	1.0	100	5.360	60.5	48.9	60.9	53.8	35.4	32.5	55.2	49.60
232	11	1.0	67	3.567	52.2	53.9	62.0	56.6	43.7	42.5	53.0	51.97
233	11	1.0	50	2.676	51.4	49.2	72.1	62.3	41.7	48.2	55.0	54.27
234	11	1.0	33	1.781	53.5	56.6	72.0	69.8	50.1	41.8	63.7	58.21
241	11	2.0	100	4.520	71.7	63.9	62.4	59.4	49.2	48.6	63.0	59.75
242	11	2.0	67	2.968	69.9	59.7	71.9	62.9	49.9	53.2	59.4	60.99
243	11	2.0	50	2.226	68.2	58.0	69.3	61.7	44.4	42.1	63.1	58.14
244	11	2.0	33	1.481	67.5	67.9	74.4	69.2	60.1	44.3	70.9	64.89
321	9	0.5	100	5.630	66.8	58.8	58.5	53.7	46.4	57.0	52.5	56.24
322	9	0.5	67	3.747	68.4	53.9	68.4	62.3	59.1	57.6	50.3	60.01
323	9	0.5	50	2.810	67.0	57.0	74.1	61.1	52.6	64.4	56.5	61.82
324	9	0.5	33	1.871	70.3	64.6	75.0	70.9	66.4	69.9	65.7	68.95
331	9	1.0	100	4.760	72.8	61.4	59.5	57.0	51.1	57.9	58.5	59.75
332	9	1.0	67	3.168	61.7	59.6	66.4	60.6	52.9	61.5	54.7	59.63
333	9	1.0	50	2.376	74.5	62.2	69.4	59.0	56.5	59.7	61.2	63.22
334	9	1.0	33	1.581	69.9	68.8	76.2	68.9	69.6	69.6	73.9	70.97
341	9	2.0	100	3.960	76.1	73.4	67.6	60.7	57.2	63.6	60.3	65.56
342	9	2.0	67	2.634	75.4	72.1	70.0	67.8	56.6	64.0	61.1	66.72
343	9	2.0	50	1.976	72.1	74.7	72.7	69.9	57.0	63.4	63.5	67.62
344	9	2.0	33	1.315	71.4	75.3	74.3	68.1	71.6	64.4	70.6	70.83
421	8	0.5	100	5.160	79.0	59.9	56.9	56.7	63.9	76.2	54.6	63.86
422	8	0.5	67	3.434	80.4	68.7	64.4	62.7	66.6	75.7	55.7	67.76
423	8	0.5	50	2.534	79.3	65.9	71.8	63.4	68.4	74.4	62.7	69.41
424	8	0.5	33	1.715	81.6	69.9	70.0	71.7	74.7	76.9	66.9	73.07
431	8	1.0	100	4.461	77.9	63.5	61.1	56.2	63.9	69.4	59.4	64.48
432	8	1.0	67	2.968	76.6	67.0	68.3	63.8	66.0	78.0	53.0	67.52
433	8	1.0	50	2.226	76.0	61.7	69.9	62.7	65.6	76.9	59.0	67.38
434	8	1.0	33	1.481	80.0	72.9	76.2	70.7	75.6	77.4	71.7	74.92
441	8	2.0	100	3.691	81.4	64.0	69.2	66.9	67.0	75.6	64.7	69.85
442	8	2.0	67	2.456	80.4	72.5	71.7	68.9	65.6	77.9	60.7	71.10
443	8	2.0	50	1.840	78.2	66.9	74.0	68.9	68.6	77.8	71.0	72.19
444	8	2.0	33	1.225	78.0	71.1	76.9	71.9	79.5	82.6	69.4	75.63

Table 7.1 System ID's and parameters, together with mean degradation rating on each of the seven test sentences (see heading), and all seven sentences pooled. (See text for more details.)

speaker or other sentence with about the same frequency. No system and no sentence followed itself.

In addition to counterbalancing the presentation sequence, we tried to further reduce sequence effects, and thus improve the reliability of the data, by fading in and out a continuous speech babble at the same level as the speech, during each inter-stimulus interval. (This method is described further below, in Section 7.4.4.) Seven experimental tapes were recorded. Stimuli were presented in blocks of ten, at a rate of one every 7.5 seconds, with a longer gap between blocks. The subject's task was to rate the degradation of the stimuli he heard. This negative attribute was chosen for scaling, because the scale has a natural origin, or zero, corresponding to undegraded speech. Degradation ratings ranged between 0 and 100, with small numbers corresponding to high quality, and large numbers to poor quality. Nine normal hearing subjects served in the experiment. All of the subjects made the first two passes through the 350 stimuli, and three of them made a further three passes each.

Results

First, to check on the reliability of the data, the responses collected on each pair of passes through the 350 stimuli were correlated, for each subject. All correlations were significant, all but three well beyond $P<.001$. Therefore,

although there was some variability between subjects, all the subjects apparently gave highly reliable data.

The mean degradation rating was calculated for each system, both by sentence, and pooled across all seven sentences. The mean ratings are shown in Table 7.1, and the pooled means are plotted in Fig. 7.3. Each system is identified by three digits, corresponding to its parameter level for P, Q, and R, respectively. Thus system 231 used level 2 of P (11 poles), level 3 of Q (1.0 dB) and level 1 of R (100 fps), as shown in the key to the figure. The 110 kbps PCM speech, used as undegraded anchor, is labelled "000." The mean ratings (N.B. not the raw ratings) have standard deviations ranging between 1.0 and 1.7 degradation points. Any difference between two plotted means that is larger than about 4-5 points is likely to be significant at $P<0.05$, and some much smaller differences were also significant. (The results of t-tests between each pair of systems are described in [19].)

Fig. 7.3 shows the effects on degradation of decreasing bit rate by: a) reducing the number of poles (top); b) coarsening the quantization step size (middle); and c) decreasing the frame rate (bottom). In each case, the two remaining parameters are held constant: each line represents a family of vocoders that differ in only one parameter. Comparing the slopes of the lines in the three parts of the figure shows dramatically that reducing

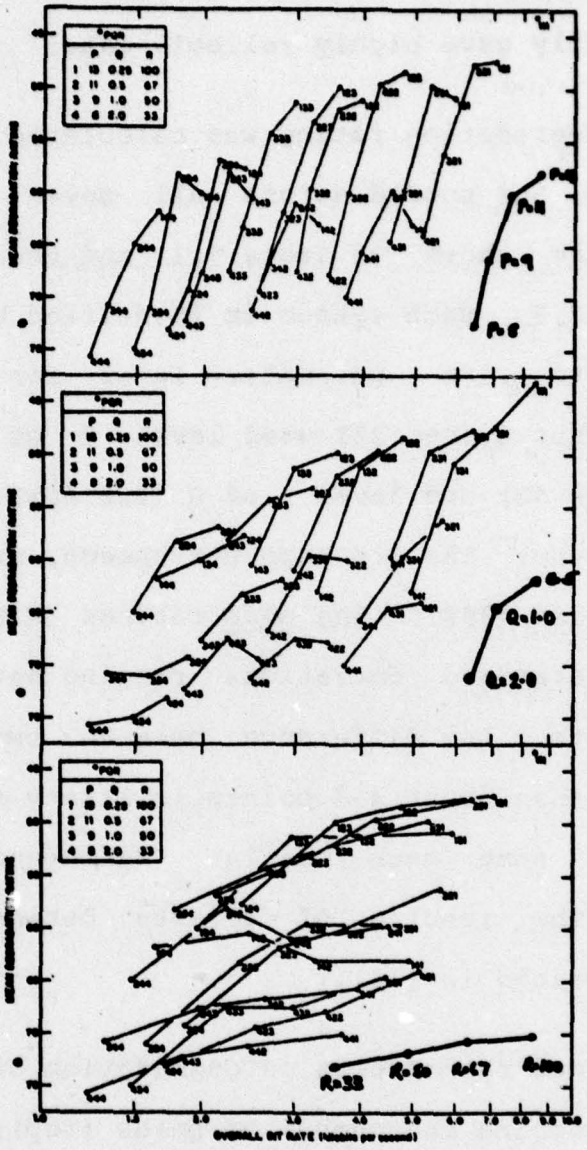


Fig. 7.3 Degradation rating vs. overall bit rate for 49 vocoders. The effect on degradation of changing the number of poles (top panel); the quantization step size (middle panel); and the frame rate (bottom panel). See text for details.

the frame rate (Fig. 7.3, bottom) yields the largest savings of bit rate for the smallest loss of quality, and that for many of the systems the loss of quality shows no knee, even at the lowest frame rate. The flatness of these lines justifies our enthusiasm for variable frame rate systems, whose superiority we document further in later sections.

Secondly, inspection of Fig. 7.3, top, shows that the rate of quality loss per bit saved is most severe for savings gained by reducing the number of poles. There is a sharp knee in most of the functions at 11 poles -- it is unfortunate that we did not also include 10 poles, although our other work suggests that 11 poles is in fact the lowest number that yields good quality with male voices, with a 5 kHz speech bandwidth.

7.3.2 Speech Quality Testing of Some VFR Vocoders

VFR transmission of LPC vocoder coefficients is a technique for reducing the average transmission rate without appreciable loss of quality (see Section 4). The technique transmits parameters at a variable rate in accordance with the changing characteristics of the speech signal. To demonstrate the soundness of the rationale for VFR transmission, an experiment was performed to compare VFR with two other methods for reducing the bit rate: (a) reducing the number of poles, and (b) increasing the quantization step size of the LAR

coefficients. The VFR scheme tested used a transmission decision based on the log likelihood ratio, with a single threshold, as described in Section 4.1. (Our superior perceptual-model based scheme, whose testing is described in Section 7.3.3, was a later development.) Thirty-two stimulus sentences were prepared by passing four utterances (2 sentences x 2 speakers) through eight vocoder systems. The vocoders were specified by a 2 x 2 x 2 factorial design; two values were assigned to each of the three parameters: average frame rate, number of poles, and quantization step size. Eight listeners made 7-point category ratings of quality degradation. The results of the experiment show that, of the three methods studied, the VFR technique produced the highest quality at any given transmission rate (or, equivalently, yielded the lowest bit rate for a fixed level of speech quality). The results of this study have been published, and the published paper is reproduced as Appendix 10.

The present study had the explicit aim of comparing systems that differed along three dimensions. We adopted a factorial design, in which two values of each of the three parameters occurred in every possible combination. The resulting systems produce a wide range of qualities. Each system used either 11 or 8 poles. The LAR coefficients were quantized in steps of either 0.5 dB or 2.0 dB. LPC analysis of the speech signal was carried out at 50 fps, and the log likelihood ratio threshold of the VFR

scheme was set to either zero dB, in which case every analyzed frame was transmitted, yielding a fixed frame rate of 50 per second, or 2.5 dB, which resulted in a variable frame rate that averaged 23.3 per second. Note that 2.5 dB represents a very coarse threshold, and that the resulting average frame rate is less than 60% of the average frame rate of the VFR systems in the study reported above (Section 7.2). Pitch and gain were coded in 6 and 5 bits respectively, and transmitted at a constant rate of 50 fps for all 8 vocoder systems.

A subset of the thirty-six test sentences used in the first study was selected. To ensure that the subset was representative of the whole set of 36, we chose the two "general" sentences (i.e. Nos 5 and 6), since between them these contain most of the English phonemes. Two speakers were then selected, one male and one female, such that the vectors corresponding to their productions of the two general sentences were separated as widely as possible in the MDPREF solution space of the earlier study. To confirm that these four stimulus sentences were adequately representative, we repeated the MDPREF analysis of the earlier study, using only the subset of data collected on the four sentences. The solution obtained was similar to the solution obtained with the whole set of 36 sentences, and achieved the same orthogonal separation of the systems by number of poles, and by frame rate. (Canonical correlations between the first three

linear composites for the two solutions were 0.978, 0.915, and 0.428.) This test confirmed that the selected subset was indeed representative.

The four sentences were passed through the eight simulated vocoders, and were recorded in two random orders on the stimulus tape, with order of sequential presentation counterbalanced fully across system pairs, and as far as possible across sentence pairs, with the constraint that no system and no sentence should follow itself. Eight subjects were then run individually through two exact repetitions of the tape -- although the subjects were not aware of the repetition. Thus each subject made four ratings on each of the 32 stimulus sentences. They rated the degradation of what they heard on a seven-point scale, 1-7, with "overflow bins" (0 and 8) at each end. That is, if a stimulus sounded appreciably better than a previous one labelled with a "1", the subject was allowed to use a "0" response.

Results

The mean ratings assigned to the eight systems are shown in Fig. 7.4, where the ratings are plotted against overall bit rate including pitch and gain. Lines join each pair of systems that differ in only a single parameter: solid lines join all pairs of systems that differ only in frame rate; dashed lines join pairs of systems that differ only in the number of poles; and dotted lines join pairs that differ only in quantization step size.

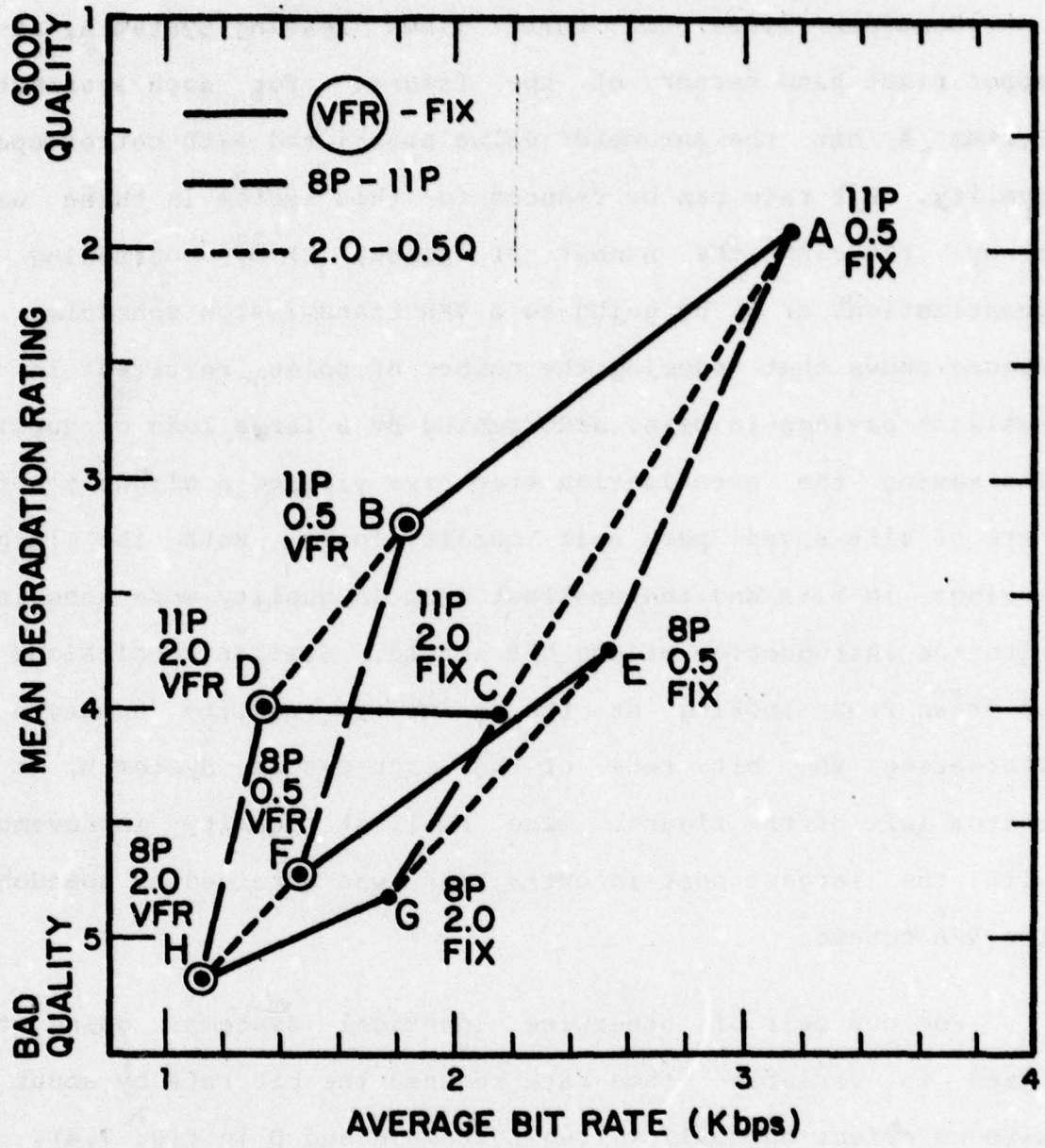


Fig. 7.4 Degradation rating vs. average bit rate for 4 fixed-rate and 4 VFR vocoders. See text for details.

Consider first the three lines leaving System A, at the upper right hand corner of the figure. For each parameter, System A has the parameter value associated with better speech quality. Bit rate can be reduced for this system in three ways: 1) by reducing the number of poles, 2) by coarsening the quantization, or 3) by going to a VFR transmission schedule. The figure shows that reducing the number of poles resulted in the smallest savings in bits, accompanied by a large loss of quality. Increasing the quantization step size yielded a slightly better rate of bits-saved per unit quality-loss. Both the largest savings in bits and the smallest drop in quality were associated with the introduction of the VFR scheme. Similar conclusions can be drawn from looking at the gains in quality achieved by increasing the bit rate of the worst system, System H, at the bottom left of the figure. The smallest quality improvement, with the largest cost in extra bits, was obtained by abandoning the VFR scheme.

For one pair of otherwise identical systems, going from fixed to variable frame rate reduced the bit rate by about 40% with no effect on quality (see Systems C and D in Fig. 7.4). All but three of the quality differences, between pairs of systems joined by lines, are extremely significant -- that is, well beyond the $P<.001$ level. The three exceptions were 1) the quality difference between Systems C and D, which was not

significant; 2) the difference between Systems G and H, which just failed to reach significance at the P<.05 level, and 3) the difference between F and H, which was just significant (P<.05).

There was a strong interaction between the speaker and the effect of number of poles. The male speaker's speech was severely degraded by the 8-pole systems, whereas the female speaker's speech was little affected. In fact, for the female speaker, reducing the number of poles yielded a rate of quality-decline per bit-saved no greater than that obtained by adopting VFR transmission. The relative speech quality of systems using 13, 11, and 9 poles on a particular sentence was highly dependent on the mean fundamental frequency in the test sentence. It is likely that the critical variable is not the fundamental frequency, but rather the length of the speaker's vocal tract, which tends to correlate highly with fundamental (large men have low voices).

Conclusions

Our results confirm that VFR transmission can yield substantial savings in bit rate, with only minor loss of quality. The rate of bits saved, per unit quality loss, is highest for savings achieved by VFR transmission, and lowest for those achieved by reducing the number of poles used in spectral modelling -- at least for the parameter values studied here. Secondly, there are major interactions between perceived speech

quality and the fundamental frequency of the talker, for some systems.

7.3.3 Quality Testing of a Perceptual-Model-Based VFR System

Subjects judged the degradation of quality caused by processing speech through six LPC vocoder systems. Two of these systems were versions of our new VFR system, based on a perceptual model (PM) of speech (cf. Section 4.2). The third was our earlier log-likelihood-ratio VFR system, and the remaining three were fixed-rate systems, one with a frame rate of 33 fps, roughly equal to the average frame rate of the PM systems, another with a frame rate of 100 fps, equal to the peak rate of the PM systems, and a third that had an intermediate rate of 50 fps. Stimulus materials were the six phoneme-specific sentences read by each of six speakers, three male and three female, as described in Section 7.2.1. The results show that the quality of the PM systems equalled or surpassed that of the 100 fps fixed-rate system, at about one third of the bit rate.

Since we have demonstrated the correctness of the rationale underlying VFR transmission (Section 7.3.2), the next question to address is whether a better strategy can be developed for deciding which frames of speech data should be transmitted. Such an improved strategy, based on a perceptual model of speech, was described above in Section 4.2. The purpose of the present study

was to make a formal comparison of subjective speech quality between (1) our improved VFR scheme (2 versions), (2) our earlier log-likelihood-ratio VFR scheme, and (3) three related fixed-rate systems. All six systems included in the test transmitted different subsets of the spectral, pitch and gain data which resulted from analyzing the input speech at a rate of 100 fps, using an 11th order predictor. Each frame of spectral data was coded in 46 bits: 6 bits were allocated to the first LAR; 5 bits each to the second and third; 4 bits each to the fourth through ninth; and 3 bits each to the tenth and eleventh LARs. Pitch and gain were coded in 6 and 5 bits respectively. Average bit-rate and frame-rate data for each of the six systems included in the test are shown below.

I.D.	BPS	Frames per second		
		LARs	Pitch	Gain
Fixed Rate:				
F100	5700	100	100	100
F50	2850	50	50	50
F33	1900	33	33	33
Variable Rate:				
VFR-1	2320	36	34	28
PML	1880	27	34	40
PMH	2120	31	34	40

Table 7.2 Overall bit rates, and frame rates for Coefficients (LARs), Pitch, and Gain, for the three fixed-rate and three VFR systems tested.

The first fixed rate system, labelled F100, transmitted at 100 fps -- that is, every frame of data analyzed was also transmitted. The overall bit rate of the F100 system was 5700

bps (46 spectral bits + 11 pitch-and-gain bits, 100 times per second). The other two fixed rate systems, labelled F50 and F33, transmitted every second and every third frame of the data analyzed at 100 fps, respectively. The F50 system was included because its bit rate and quality are comparable to those of LPC-I, specified for the ARPANET. However, F50 differs from LPC-I (a) in signal sampling rate (10 vs. 6.7 kHz); (b) in bits per frame (46 vs. 56); and (c) in the pitch extraction scheme. The third fixed-rate system, F33, was included to demonstrate the substantial degradation of quality associated with a simple fixed-rate system transmitting at about the same average bit rate as the VFR systems.

The three VFR systems represent two different transmission strategies, one using a log-likelihood ratio decision, and the other two a perceptual-model based decision. The latter two systems differ only in the thresholds for determining which frames of spectral data should be transmitted.

The log-likelihood ratio system, labelled VFR-1, selected frames of LARs (analyzed at 100 fps, as for the fixed-rate systems) using our earlier single-threshold log-likelihood ratio scheme, with the threshold set at 1.5 dB. Pitch and gain data (coded in 6 and 5 bits, as above) were selected for transmission using our double-threshold FAP scheme (cf. Section 4.3.1), applied to the quantized values. Threshold values were 0 and 1

quantized steps for pitch, and 2 and 3 quantized steps for gain.

The two perceptual-model based systems (PMH and PML) selected spectral frames for transmission, from the same data analyzed at 100 fps, using the simplified VFR scheme described in detail in Section 4.2.8. The system labelled PML (Lower rate) used a threshold of 1.3, whereas PMH (Higher rate) used a threshold of 1.0. Both systems transmitted exactly the same pitch and gain data. The quantized pitch data were selected for transmission by the single-threshold FIT scheme (Section 4.3.2), with a threshold of 0 steps, and the quantized gain data were selected by the double-threshold FIT scheme, with thresholds of 0 and 1 steps.

The speech materials consisted of 36 test sentences: the set of six phoneme-specific sentences read by the six speakers described above. Each of the 36 test sentences was processed by each of the 6 LPC systems, yielding a total of 216 stimulus sentences. These were recorded on tape in two separate orders, each counterbalanced so that each speaker followed each other speaker an equal number of times, and similarly for the sentences and systems.

After some preliminary practice, subjects rated the subjective quality of each of the 216 stimulus sentences on an 8-point category scale, with "overflow bins" of 0 and 9. Each

tape was rated in a separate separate 25-minute session. The subjects were instructed to use the full range of the rating scale, assigning 8's to the "best" quality stimuli, and 1's to the "worst." The overflow bins were to be used only when an extreme rating (1 or 8) had been assigned to the previous stimulus, and the following stimulus seemed to be even more extreme. The five subjects who served were all highly familiar with vocoded speech.

In Fig. 7.5, the mean ratings across all speakers, sentences, subjects, and replications (the 2 sessions for each subject), are plotted against mean overall bit rate, for each of the six systems. The points representing the three fixed rate systems are joined by one line, and those for the three VFR systems by a second line. For the fixed rate systems, reducing the frame rate from 100 fps to 50 fps resulted in a slight gain in quality, but further reducing it to 33 fps produced a major loss of quality. The three VFR systems apparently produced quite similar quality, roughly equivalent to the F100 and F50 systems, but at a bit rate comparable to the F33 system.

T-tests showed that several of the apparently quite small differences in quality were highly reliable. The ratings were converted to differences for the purpose of the t-tests: the variate tested was the difference in rating assigned to the two systems being compared, for the <u>same</u> sentence by the <u>same</u> speaker

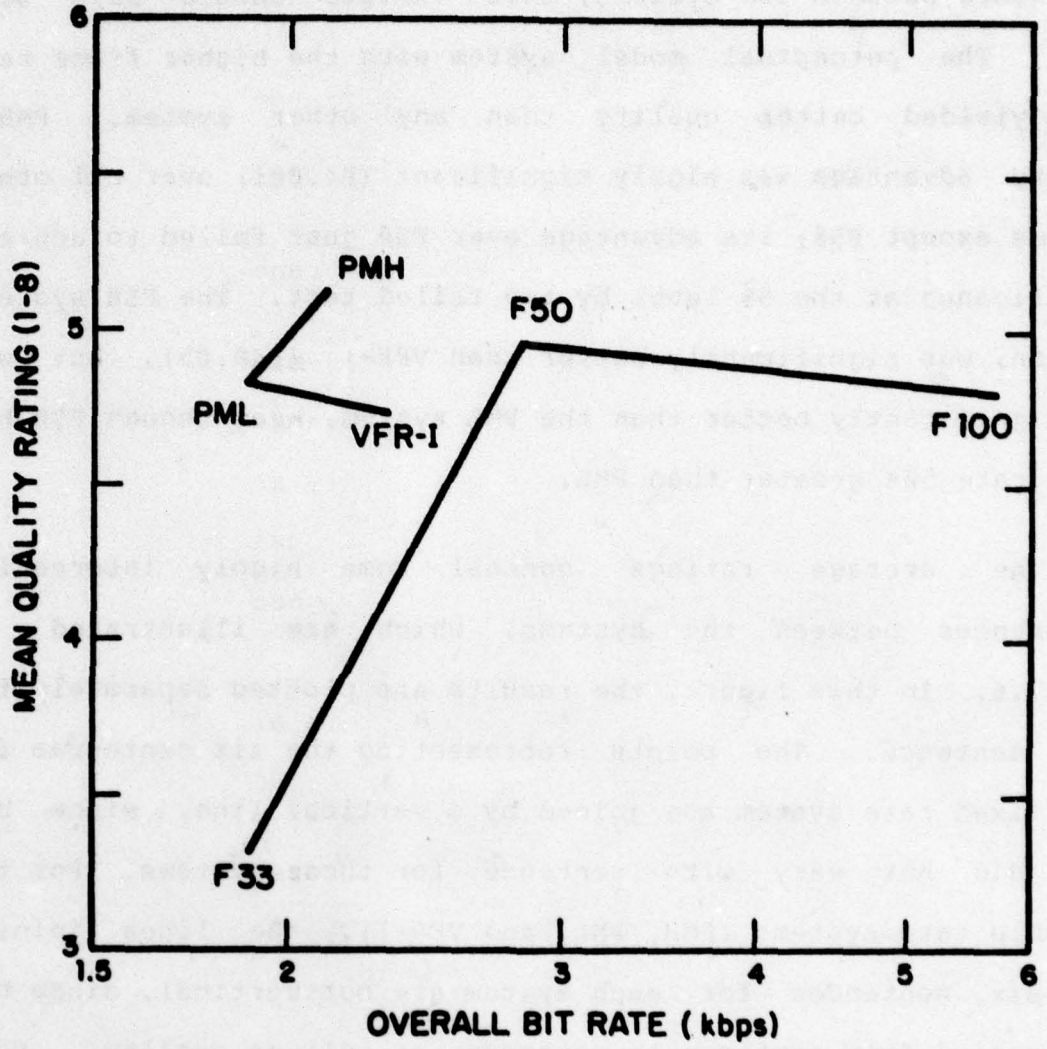


Fig. 7.5 Mean quality rating vs. overall bit rate for 3 fixed-rate and 3 VFR systems, all transmitting frames from the same data base. See text for details.

judged by the same subject in the same session. If there was no difference between two systems, this variate should have zero mean. The perceptual model system with the higher frame rate (PMH) yielded better quality than any other system. PMH's quality advantage was highly significant ($P<.001$) over all other systems except F50; its advantage over F50 just failed to achieve significance at the 5% level by two tailed test. The F50 system, in turn, was significantly better than VFR-1 ($P<0.05$), but was not significantly better than the PML system, even though F50 had a bit rate 50% greater than PML.

The average ratings conceal some highly interesting differences between the systems, which are illustrated in Fig. 7.6. In this figure, the results are plotted separately for each sentence. The points representing the six sentences for each fixed rate system are joined by a vertical line, since bit rate did not vary with sentence for these systems. For the variable rate systems (PMH, PML, and VFR-1), the lines joining the six sentences for each system are not vertical, since bit rate varied from sentence to sentence, as well as quality. Each data point consists of a digit, keyed in the caption, which identifies the sentence used.

When processed by the PMH system, each of the six sentences obtained an above-average rating, and the PMH system was never significantly outperformed by any other system, on any sentence.

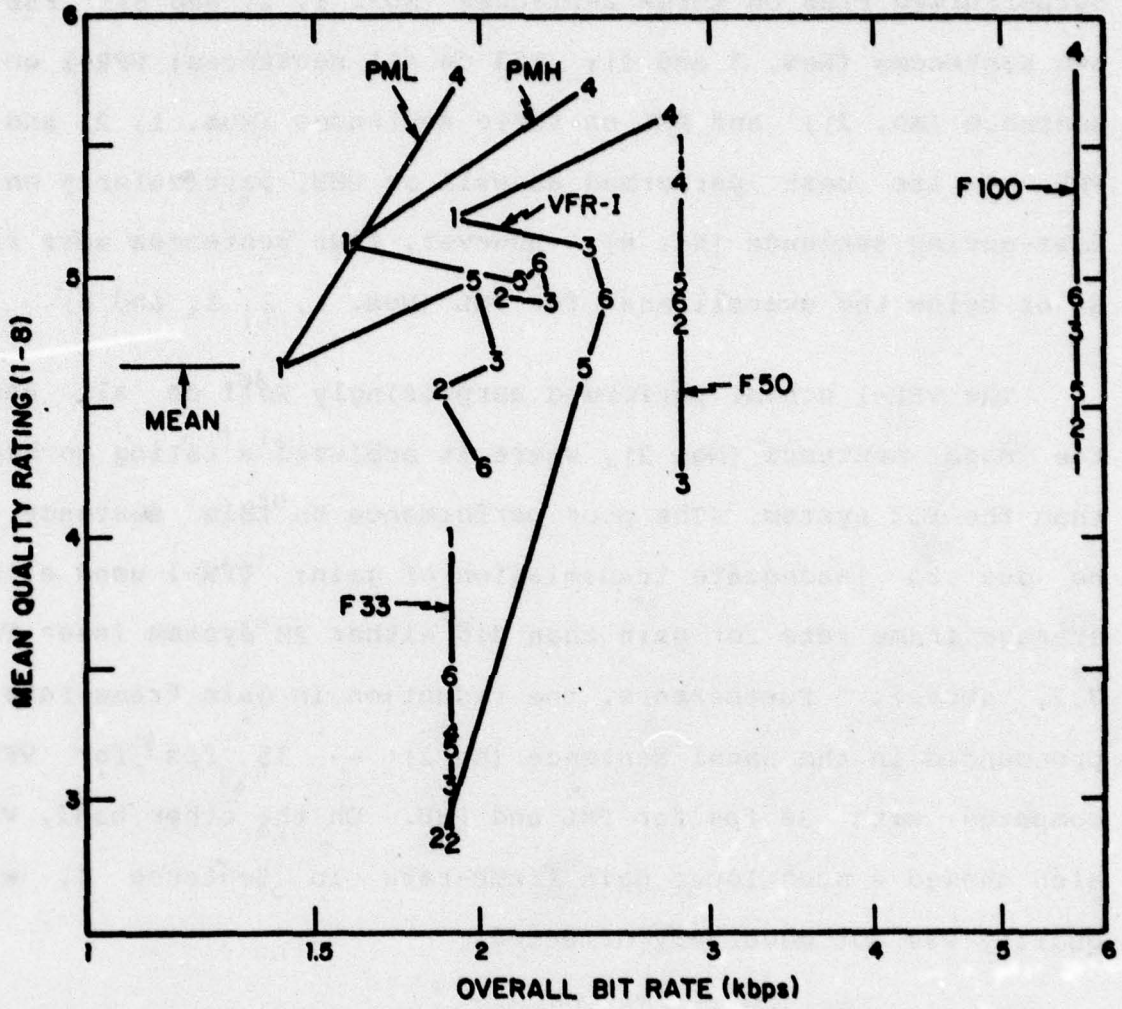


Fig. 7.6 Mean quality rating vs. overall bit rate for 3 fixed-rate (F33,F50,F100) and 3 VFR (VFR-1, PMH, PML) systems. Each of the six digits joined by a line to represent a system's performance correspond to the system's performance on a particular test sentence, as follows:

1. Why were you away a year, Roy?
2. Nanny may know my meaning.
3. His vicious father has seizures.
4. Which tea-party did Baker go to?
5. The little blankets lay around on the floor.
6. The trouble with swimming is that you can drown.

On the other hand, t-tests showed that PMH significantly outperformed F100 on three sentences (Nos. 1, 2, and 5); F50 on two sentences (Nos. 3 and 4); F33 on all sentences; VFR-1 on one sentence (No. 2); and PML on three sentences (Nos. 1, 2, and 6). PML at its best performed as well as PMH, particularly on the fast-moving sentence (No. 4). However, four sentences were rated at or below the overall mean for PML (Nos. 1, 2, 3, and 6).

The VFR-1 scheme performed surprisingly well on all except the nasal sentence (No. 2), where it achieved a rating no higher than the F33 system. The poor performance on this sentence may be due to inadequate transmission of gain: VFR-1 used a lower average frame rate for gain than did either PM system (see Table 7.2, above). Furthermore, the reduction in gain frame-rate was pronounced in the nasal sentence (No 2) -- 15 fps for VFR-1, compared with 30 fps for PML and PMH. On the other hand, VFR-1 also showed a much lower gain frame-rate in Sentence 1, whose quality was not adversely affected.

The F100 system performed surprisingly badly on Sentences 1 and 2, both of which have continuous voicing and no very large or rapid changes of spectrum. Earlier work showed that these two sentences were particularly sensitive to distortions introduced by too coarse quantization. The "wobbly" quality of these two sentences, as processed by the F100 system, may be due to instability as the quantization levels are slowly swept, in the

absence of other spectral discontinuities in the speech material. The results of such instability would be more noticeable at 100 fps than at 50 fps, both because the instability would have more opportunity to occur, and also because the periodicity of the resulting distortion would be nearer to that of the voice fundamental.

Conclusions

1. The Perceptual Model scheme yielded the same or even better quality than the fixed rate scheme on which it was based, and at substantially lower bit rates.

2. The PMH system appears to have achieved a point of diminishing returns: reducing the coefficient frame rate from 31 fps to 27 fps (in PML) yielded significantly worse quality on three of the test sentences, with insignificant savings in bit rate.

3. Since the PMH system equalled or surpassed the F100 system on which it was based, further improvements in quality can be obtained only by improving the design decisions that went into the F100 system. Several subsequent developments have suggested possible improvements.

4. The phoneme specific sentence material yields results that have high diagnostic value: for example, the poor performance

of VFR-1 on nasals might never have been verified if homogeneous testing materials had been used.

7.4 Miscellaneous Topics

7.4.1 Phoneme-Specific Intelligibility Test

We tried out a phoneme-specific intelligibility test slightly modified from one that was developed by Stevens [20,21]. The test has two parts, one for consonants and one for vowels. It is a nonsense-syllable test, using closed response sets of 4-8 items. Both of these factors increase the difficulty of the test over that of the DRT [22], which is the only other test available with similar diagnostic power. Weaknesses of the DRT are that it tests only single consonants in initial position, and the response set for each item contains only two English monosyllables, whose initial consonants are a minimal pair, differing in only one distinctive feature. The small response set greatly reduces the efficiency of the test, since chance performance is 50%. In contrast, the Phoneme-Specific Intelligibility test covers vowels, and single and clusters of consonants both in pre-stress and in final position. The stimulus items are nonsense syllables of the form /ə'C1VC2/, where /ə/ is an unstressed schwa like the first syllable of "about," C1 and C2 are consonants, and V is a stressed vowel.

The complete test consists of 14 separate subtests. The first ten are consonant tests, each of which uses a single closed set of consonants from which C1 and C2 are drawn. There are four versions of each consonant subtest, two of which use one pair of vowels as syllable nuclei, and two using a second pair of vowels. A typical consonant test list is shown in Fig. 7.7. Each consonant in the closed response set appears four times in each list, once preceding and once following each of the context vowels. In addition, there are three unscored filler items (ringed numbers in the figure) added to prevent subjects from using the symmetry of the test to aid their responding. The vowel tests are similar, except that each vowel appears four times in each list, in symmetrical consonant context, and there are three different sets of consonant contexts for each vowel subtest. The complete set of 64 lists is given in Appendix 11. The test is in most respects identical with that reported by K. N. Stevens [20,21]. The complete test has never been published before, and we thank Prof. Stevens for permission to include it here.

One male and one female talker each recorded half of the 64 test lists. We ran preliminary tests on a small subset of the lists, using four simulated vocoders from those specified for the test of our quality assessment method, described above in Section 7.2.3. Although the test results were quite encouraging, we

TEST NO. 5AM NAME _____ DATE _____

CONSONANTS: b d m n v z
VOWELS: æ ʌ

1. v ʌ m
2. n ʌ z
3. d æ m
4. z ʌ d
5. d ʌ n
6. z æ b
7. n ʌ b
8. m æ n
9. v æ m
10. b æ z
11. n æ v
12. b ʌ v
13. m ʌ z
14. b ʌ n
15. d æ d

Fig. 7.7 A representative consonant test list from the Phoneme Specific Intelligibility Test (Stevens, 1962). The whole test is given in Appendix 11.

abandoned further testing, since processing the test lists through simulated, as opposed to real-time, vocoders was prohibitively time consuming. More details of the pilot tests can be found in [23].

The test is probably the best available for generating high-quality diagnostic data about real-time systems, but even here it has two drawbacks. The test is long, taking several hours for each subject, for each tested system. Secondly, some of the lists require the listeners to be familiar with phonetic symbols, which means that additional training is necessary if skilled subjects are not available. A further problem is that the diagnostic data consist of the pattern of errors made, and if the systems under test are highly intelligible it may be necessary to run large numbers of subjects, or repeat lists, to accumulate sufficient errors. Of course, other diagnostic tests suffer the same disadvantage, especially the DRT which forces a choice between only 2 alternatives for each test item, resulting in a high chance performance level. An alternative method for increasing the number of errors is to degrade the acoustic (or other) environment of the speaker or listeners. This procedure is appropriate only if the added degradation remains within the range to be expected in the final application.

The high face-validity of the test procedures, together with their potential for diagnosing problems with specific types of

phonemes, make the foregoing drawbacks acceptable for testing real-time systems, although they are probably more appropriate to the development of improved vocoding systems than to routine acceptance testing.

7.4.2 Effects of Lost Packets on Intelligibility

Decisions on how much speech to encode in one packet for transmission over the ARPANET (and for Packet Radio) have been made on the basis of two factors: overhead, and delay. Each packet contains a fixed number of header bits, etc., and the cost of this overhead decreases as more speech is encoded in a packet. On the other hand, packetizing speech introduces a delay equal to the duration of a packet's contents (in addition to other delays due to path length and network response). Delays have serious disrupting effects on conversations [24], so delays must be minimized.

NSC Note No. 78 [25] was written to point out that there is a further factor that should be considered in deciding how much speech to encode in a packet: the effect on intelligibility of lost or delayed packets. Work with interrupted speech, and with speech alternated between the ears, and with "temporally segmented" speech (summarized in [26]) shows that silent intervals inserted into continuous speech, whether the silence displaces or delays the speech waveform, have a maximally

disruptive effect on intelligibility when the silent intervals are in the range 100 to 300 ms. This is exactly the range of silent intervals that would be introduced into speech if reconstruction of the speech had to continue in the absence of a packet, either lost or delayed. An alternative to leaving a silent interval is to repeat the preceding packet, but this may introduce intelligibility problems of its own.

A possible solution was suggested, that involved interleaving the successive frames of speech data in two independent packets, one containing even-numbered frames and the other odd-numbered frames. A lost packet would then result in a brief burst of interrupted speech, with silent intervals of 20 ms, which would have no effect on intelligibility. The cost would be increased delay. More details can be found in [25].

7.4.3 Descriptor Inventory for Subjective Quality

A listening test was conducted to identify terms descriptive of vocoded speech for listeners unfamiliar with vocoding techniques. The test was carried out in two stages. In the first stage, the listeners were requested to list adjectives or phrases that they considered descriptive of the speech to which they were listening. In the second stage, they were provided with lists of words and phrases, and asked to judge the appropriateness of each of the items on the lists to the speech.

The speech samples were those generated for the experiment described in Section 7.3.2, together with a 110 kbps PCM version of each of the four sentences, to act as undegraded anchor. Sentences were heard in pairs. The first member of a pair was always the unprocessed PCM version of the sentence; the second member was one of the eight processed versions of the same sentence spoken by the same talker. Listeners were encouraged to attend to the ways in which the standard (unprocessed) and test (processed) samples differed.

Listeners were 17 undergraduates who reported normal hearing, and had no previous experience with vocoded speech. First, subjects listened to several items and then began making their list of descriptors. After 10 minutes, these lists were gathered, and previously prepared check lists were distributed. The listeners rated each of the words and phrases on these lists, on a 10-point scale, for its appropriateness as a descriptor of the processed speech they were hearing. Meanwhile, another list was composed consisting of items produced by the listeners during the first stage, and the listeners continued the test by assigning scale values to these new terms.

Table 7.3 shows the 127 descriptors presented for rating during stage 2 of the test. Table 7.4 shows the ten words that received the highest ratings, considering all of the listeners, and also considering two subsets of "best" listeners. "Best" was

In some of the pairs, the second sentence has a _____quality.

1___blary
2___boomy
3___bouncy
4___brassy
5___breathy
6___burbly
7___buzzy
8___chirpy
9___choppy
10___chattery
11___clean
12___clicky
13___clipped
14___coarse
15___computer-like
16___crackly
17___creaky
18___crisp
19___croaky
20___damped
21___dead
22___deep
23___diffused
24___disconnected
25___distinct
26___distorted
27___drone-like
28___dull
29___eddying
30___electronic
31___even
32___frizzy
33___flat
34___fluctuating
35___fuzzy
36___garbled
37___grating
38___grinding
39___gruff
40___gurgly
41___guttural
42___hissy
43___hollow
44___human
45___hum-like
46___hushed
47___husky
48___indistinct
49___jangling
50___jerky
51___mellow
52___metally
53___monotone
54___murmury
55___musical
56___muted
57___nasal
58___natural
59___noisy
60___oscillating
61___piercing
62___hi-pitched
63___pulsating
64___pure
65___raspy
66___reed-like
67___regular
68___resonant
69___reverberant
70___rich
71___ringy
72___rough
73___scratchy
74___sharp
75___sharp-edged
76___shivery
77___shrill
78___silvery
79___slurred
80___smooth
81___smooth-edged
82___soft
83___spitty
84___spluttery
85___sputtery
86___squawky
87___squeaky
88___steady
89___stifled
90___strained
91___strident
92___subdued
93___telephonic
94___throbbing
95___tinny
96___trill
97___twangy
98___tweeting
99___twittery
100___unbroken
101___unclean
102___undulatory
103___uneven
104___vibrant
105___warbly
106___wavery
107___wheezy
108___whirring
109___whispery
110___wobbling
111___yodelling
112___whistling
113___tinkling
114___thin
115___swishing
116___screeching
117___rumbling
118___rippling
119___radio-static
120___quavering
121___harsh
122___full
123___fluttering
124___flat
125___echoing
126___clear
127___broken

Table 7.3 Descriptor inventory

All	Best 9	Best 3
nasal	nasal	nasal
muffled	muffled	muffled
distorted	distorted	fuzzy
monotone	head cold	distorted
blanketed	garbled	stuffed up
fuzzy	dull	muted
head cold	monotone	blanketed
dull	blanketed	head cold
garbled	fuzzy	damped
muted	slurred	parrot-like

Table 7.4 Descriptors with highest utility for a) all subjects, b) the 9 most consistent subjects, and c) the 3 most consistent subjects. (See text for details.)

defined in terms of the similarity of a listener's ratings to the group mean ratings.

7.4.4 Reducing Sequence Effects in Quality Assessment

In tests of intelligibility, there is an objectively correct answer for each test item, whereas in tests of speech quality, the responses are judgments for which there is no correct answer. Consequently, results obtained in speech quality tests tend to be highly subject to context effects. The rating assigned by a subject to a particular test item depends not only on the test item itself, but on the range of qualities associated with the other systems under test, and also on which of these other systems were presented for judgment as the preceding two or three stimuli. That is, different ratings are given to a single system, depending on which system was presented on the preceding trial(s).

The usual method of combatting sequential effects is to counterbalance the presentation sequence, so that every stimulus is preceded equally often by each of the other stimuli in the set, so that biasses cancel out. Where large numbers of systems are being compared, this procedure rapidly becomes impractical since the required number of stimulus presentations increases with the square of the number of systems being compared. In the PARM test, developed by Voiers for DCA [27], the number of

stimulus presentations was kept small by comparing only six systems at a time, two of which were anchor systems that appeared in every sextet to provide a baseline for comparing different sextets. However, as Voiers points out, even these carefully devised conditions failed to adequately control the sequence effects.

Sequence effects must depend on memory of the perceived quality of the stimuli presented earlier. If the memory could be erased, the sequence effects would disappear. One possible method is suggested by recent work on auditory short term memory, on the so called suffix effect [28,29]. These results show that, when a list of items is presented for immediate recall, adding an extra item to the end of the list (the redundant suffix) interferes with the auditory memory traces of the last items in the list, even though the subjects knew what the extra item would be. That is, presenting a redundant suffix erases, at least partially, the memory traces of earlier items. Since this is precisely the effect we would like to achieve to reduce sequence effects in quality tests, we carried out a study in which we adapted the suffix effect paradigm for this purpose.

The method adopted was to fill the silent intervals between successive stimuli with speech babble. The babble consisted of a carefully controlled mix of six different voices, reading a variety of passages, which had been developed at BBN as part of a

separate project [30]. To test the method, we repeated the earlier quality study of VFR vocoders reported above in Section 7.3.2, using seven of the eight original subjects. A new stimulus tape was prepared of the same stimuli, in the same presentation order. The babble, at the same level as the signal, was automatically faded out and in again one second before and after each stimulus presentation.

Each of the two experiments showed a highly significant assimilative sequence effect. The hoped-for difference between the two experiments, ascribable to the intervening babble, was not significant by t-test ($p<0.15$), although the difference was in the desired direction, suggesting the babble may have reduced the sequence effect slightly. In support of this, all subjects reported that the task seemed easier with the babble, and that the babble made it harder to compare a stimulus with its predecessor.

Comparison of the data collected with and without babble showed that both experiments yielded highly similar results, except that the speech appeared slightly more degraded <u>with</u> babble, perhaps because the babble consisted of a mixture of voices recorded under good conditions, and may therefore have acted as an undegraded anchor against which the eight vocoder systems appeared more degraded than in the absence of the babble.

8. OBJECTIVE SPEECH QUALITY EVALUATION

Quality assessment of vocoded speech is often performed to determine the user acceptance of a vocoder, or to compare the performance of competing vocoder types, or to evaluate the different choices of a given vocoder's design parameters. Procedures used for speech quality measurement are either subjective or objective, depending upon whether or not they make use of subjective judgments from human listeners. Subjective procedures require extensive testing with human listeners, which is expensive in terms of both time and money. On the other hand, objective measures would enable evaluation to be done by computer as well as ensure uniformity in speech quality evaluation. Also, objective measures can be incorporated into the design of better quality vocoders. Of course, the validity of any objective procedure must first be established by comparing its results against subjective judgments.

Major achievements of our objective speech quality evaluation work have been: 1) Formulation of a general framework, and (2) Development of several usable objective quality measures which produce results highly correlated with subjective judgments. The results of our work have been presented in three papers, which are included in this report as Appendices 12-14. Below, we provide a brief summary of these results.

8.1 A General Framework

We formulated a general framework for the objective evaluation of vocoder speech quality, based on the following reasonable assumptions (For more details, see Appendix 12):

(1) Speech synthesized from unquantized LPC parameters (14th order LPC filter, for a speech bandwidth of 5 kHz), extracted every 10 ms, is of very good quality, compared to the original speech.

(2) Except for pitch and gain, the fidelity of the short-time speech spectrum is the principal determiner of quality.

(3) The spectrum is uniquely defined by the linear prediction filter parameters.

The first assumption gives us an anchor point, defined in terms of the unquantized LPC parameters, against which to compare quantized realizations of the same utterance. The second and third assumptions relate the filter parameters to speech quality. In this framework, then, the problem of objective quality evaluation is reduced to the following two steps: 1) For each 10 ms frame, compute an objective error as the distance or deviation between the spectrum corresponding to the unquantized LPC parameters and the spectrum corresponding to the quantized and interpolated LPC parameters; and 2) Combine all the frame errors thus computed within a speech utterance into one number, which

becomes the objective speech quality score. Notice that the described objective quality measurement procedure can be carried out when the LPC vocoder is in operation.

8.2 Spectral Distance Measures

To perform the task of step (1) above, we developed several spectral distance measures which produced results consistent with published subjective perceptual results on formant frequency difference limens. A detailed description of these measures is given in Appendix 13. Briefly, given two smooth spectra, the distance between them is computed in three steps:

(a) Normalize the two spectra by making them have either the same geometric mean (GM normalization) or the same value at zero frequency (DC normalization);

(b) Determine the error at each frequency as the magnitude of the difference in linear spectral amplitudes of the two spectra; and

(c) Compute the (weighted) norm of this error function after weighting the error with the perceived loudness function, originally developed by S.S. Stevens for a different purpose.

We chose to study in detail the use of two distance measures, denoted below as d(GM) and d(DC), which use, respectively, GM and DC normalization. In addition, we considered two other measures, d(RMS-LOG) and d(LAR), for comparative purposes; the first of

these two measures computes the spectral distance as the rms value of the difference in the log spectral amplitudes of the two spectra, and the second measure is the Euclidean distance between the two p-vectors of LARs corresponding to the two spectra. Since LARs are readily available in the problem at hand, using the latter measure is computationally much less expensive than using any of the other three measures.

The task, in step (2) above, of combining the frame errors into one number involves first weighting the frame errors with a suitable time-weighting function to reflect the relative importance of the individual frames to perceived speech quality, and then averaging the weighted frame errors. A detailed account of the results of our work on this task, as well as the results of correlation tests between our objective quality scores and subjective judgments are given in Appendix 14. Below, we give a brief summary of these results.

8.3 Time Weighting of Frame Spectral Errors

We investigated the two time-weighting methods described below.

(i) Filter Gain Weighting: In this method, we make the reasonable assumption that frame errors in low energy regions of an utterance have a smaller influence on quality judgments than those in high energy regions. For example, even large changes in

the spectrum may not be detected by the listener if the total energy in the spectrum is low. We considered the weighting as a function of the frame speech signal energy per sample expressed in decibels. A piecewise linear weighting function was found to produce good correlation between the resulting objective scores and the corresponding subjective test results.

(ii) Weighting Based on Our Perceptual Model: In the second type of (implicit) time weighting that we explored, we employed as anchor or reference our perceptual model of speech instead of the 100 fps LPC analysis data. That is, we used the analysis data only for those frames for which our new automatic VFR scheme (see Section 4.2) decided to transmit; for all other frames, we obtained the LPC data via linear interpolation between the adjacent transmitted frames. In addition, we employed an explicit time-weighting in which frame errors for the transmitted frames are weighted with unity, while other frame errors are weighted with a fraction depending on the duration of the transmission interval to which they belong.

8.4 Time-Average of Weighted Frame Errors

There are a number of different ways of combining the weighted frame errors into one number. The simplest time-average is the arithmetic mean or straight average. We also considered a two-term composite average: the first term is simply the

arithmetic mean over the whole utterance, and the second term is the arithmetic mean over the top 10% of the frame errors. A third measure we investigated is the above composite average but with the second term computed over a variable percentage of large frame errors; this variable amount was decided by the "skewness" of the frame error distribution over the whole utterance.

8.5 Correlation with Subjective Judgments

In our initial studies, we compared our objective speech quality scores against subjective test results obtained for the five utterances JB1, AR4, JB5, RS6, and DK6, and for 22 of the 49 vocoders included in our factorial subjective speech quality study (see Section 7.3.1). We computed two types of correlation between the objective and subjective data: (1) regular, or Pearson's product-moment, correlation (we shall call this simply correlation); and (2) rank order, or Spearman's rank, correlation. For the second type, two sets of ranks are first assigned to vocoders under study using separately objective and subjective data, and then regular correlation is computed between the two sets of ranks. Correlation scores were used as a means of choosing the parameters of the time-weighting and time-averaging schemes discussed above.

Results obtained using the correlation study are briefly summarized below:

(i) Using the spectral distance measure d(DC) generally produced substantially lower correlations than using any of the other three measures investigated. Therefore, we eliminated the measure d(DC) in all our subsequent studies.

(ii) Correlation scores obtained for the utterances from male speakers were generally higher than those for the utterances from female speakers. Also, analysis of our subjective speech quality test results showed that subjective rating scores for the utterances from female speakers were relatively constant over the range of the number of poles (or LPC order) considered (9-14 poles); in contrast, the rating scores for male speakers exhibited a wide range of variation [13]. This suggested the variation of the LPC order for the anchor system as a function of the average fundamental (or pitch) of the speaker over the whole utterance. This technique was found to slightly enhance the correlation scores for the utterances AR4 and RS6.

(iii) An important achievement of our objective speech quality evaluation work has been that we obtained relatively high correlation scores. For the measure d(GM), correlation for individual utterances varied between 0.8 and 0.96; rank correlation had the range from 0.8 to 0.9. For the measure d(RMS), these ranges were found to be: 0.85 - 0.94 for correlation, and 0.83 - 0.88 for rank correlation. For the

measure d(LAR), we obtained the ranges: 0.79 - 0.93 for correlation, and 0.78 - 0.83 for rank correlation.

9. TOWARDS REAL-TIME IMPLEMENTATION

We cooperated with the other sites in the ARPA community in implementing an LPC vocoder that transmits speech over the ARPA Network in real time. Below, we first describe our work to develop a real-time speech facility at BBN, and then briefly summarize the specifications that we provided for ARPA LPC-II speech compression system.

9.1 BBN Speech Facility

Our signal processing system was designed to meet the needs of both the speech compression project and the then existing speech understanding project. It consists of the two computers, the SPS-41 and the PDP-11. The SPS-41 has a dual-port memory interface, and we installed a dual channel A/D and D/A converter system. We added an IMP11A interface to our system to provide a link to the ARPA Network.

In close cooperation with the Information Sciences Institute (ISI), we worked on an on-line loader system for the SPS-41. This consists of two parts, the Overlay Executive (EXEC) and the Automatic Reformatter (ARF). The EXEC is an SPS-41 program which loads information from the PDP-11 into the SPS-41. ARF reformats the output of the SPS-41 assembler in a way acceptable to the EXEC. It also provides a mechanism for attaching meaningful labels to SPS-41 program segments and locations.

We modified the LPC programs and support software supplied by other ARPA-sponsored sites and by SPS, Inc., to run on our configuration of the PDP-11/SPS-41 system and we developed a procedure for loading these programs from TENEX into the PDP-11. We worked towards locating and describing hardware problems in the SPS-41, which appeared to be the cause of system failures after short periods of successful operation. As part of this effort, our SPS-41 was moved back to SPS, where we had one person working full time trying to resolve these problems with the help of people from SPS. During that time, several hardware problems were detected and corrected. Subsequently, several versions of the back-to-back LPC software were successfully run for a considerable length of time.

We purchased an RT11 operating system for our PDP11/40. Upon delivery of this system, it was modified to permit the use of the existing Telefile/Century Data disc. This disc has a storage capacity of 500 Mbits and has been used for temporary storage of computer programs and sampled speech signals.

Our more recent work has proceeded in two directions: (1) To develop the PDP11/SPS41 system for use as a research tool, specifically for the acquisition, storage and playback of speech waveforms, and (2) To bring up a real-time vocoder system on the ARPANET.

The real-time acquisition and playback system operates in conjunction with another larger computer system, in this case the DEC System 20. In typical operation, the real-time system digitizes and stores an utterance. The user then has the opportunity of listening to the digitized utterance, displaying it, editing out such undesirable features as tape recorder pops, and in general, checking to see that the complete utterance had been digitized. Initial and final periods of silence are edited out in order to save storage space. Once the utterance has been edited and checked, the digitized waveform can be transmitted to the System 20, to be used in synthesis experiments involving different vocoder systems. Any synthetic utterances resulting from these experiments can be transmitted back to the real-time system, for the user to play out through the D/A converter. We have also implemented an interactive playback program on the PDP11, which allows the user to easily specify and play out any sequence of digitized speech signals. This program has been quite useful for running informal listening tests, and for conveniently and rapidly preparing audio tapes for demo purposes and for formal subjective speech quality tests.

We have developed support software for the system, including an FTP (File Transfer Protocol) program which allows us to transfer files between the real-time system and the System 20 or any other host on the ARPANET. We have handlers for the IMLAC

PDS-1 display computer, which runs as a peripheral to the PDP11. These handlers allow the IMLAC to be used as a high speed terminal on the PDP11 and at the same time support its display functions.

We have also worked closely with ISI to modify EPOS (Environment for Processing of On-line Speech), to work with our file structures. EPOS is required by the existing versions of the LPC vocoder.

9.2 Specifications for ARPA LPC-II System

We provided specifications, in the form of NSC Note No. 82 [7], for ARPA LPC-II speech compression system, an update of the earlier system LPC-I, for real-time implementation at various ARPA-sponsored sites. We had previously developed the following approaches for reducing the redundancy in the speech signal [1]:

(1) optimal parameter quantization using LARs,
(2) variable frame rate (VFR) transmission of LARs,
(3) variable order linear prediction, and
(4) Huffman coding.

We recommended only items (1) and (2) for LPC-II, in an attempt to reap maximum benefit for the least amount of effort in terms of changes to LPC-I. Our overall design objective in arriving at specifications for LPC-II was to achieve average

continuous-speech transmission rates of about 2200 bps. This bit rate should be contrasted with that of LPC-I which is about 3500 bps.

There are thus two major differences between LPC-I and LPC-II. These are: 1) LPC-II uses VFR transmission of LPC parameters, whereas LPC-I uses a fixed frame rate, and 2) use of new coding/decoding tables for transmission parameters. These new tables were obtained using

(a) uniform quantization of LARs;

(b) different step sizes for different LARs, based on their relative spectral sensitivities (see Section 3.1); and

(c) smaller ranges (i.e., minimum and maximum values) for reflection coefficients (or equivalently LARs), than were used in LPC-I. These ranges were obtained from real speech data. than were used in LPC-I.

Compared to LPC-I, VFR transmission yields a lower (average) frame rate, while new coding/decoding tables employ fewer bits per transmitted frame. Thus, both modifications contribute to lowering the average bit rate. These modifications were based on the results of our previous research [1].

Initially, we specified a procedure in which only the log area ratios were to be transmitted at variable frame rate; pitch and gain were to be transmitted essentially at a fixed rate. Later, in NSC Note 96 [8], we presented VFR transmission schemes

for pitch and gain also. Use of these schemes in LPC-II would lower the average transmission rate to about 2000 bps for continuous speech. (With the use of a silence detection algorithm, these average rates may drop to about 1000 bps or less.)

LPC-II has been implemented at CHI and ISI. Upon informally listening to speech from the vocoders LPC-I and LPC-II, provided to us by CHI, we found, as did CHI, that the speech quality of LPC-II was about the same as that of LPC-I. The listening tests also showed that there was room for improvement in speech quality of LPC-II by using a more perceptually based VFR transmission scheme for log area ratios than the likelihood ratio method employed in LPC-II (see Section 4.2). As part of the follow-on ARPA contract, we plan to implement a real-time LPC system that employs our perceptual-model-based VFR transmission scheme.

10. MISCELLANEOUS TOPICS

Two additional issues that we investigated during this project are reported in this section.

10.1 Coding of LPC Parameters Using DPCM

Differential Pulse Code Modulation or DPCM is a well-known method for quantizing signals which exhibit high correlation between successive samples. This method has been widely used for coding speech signals. Following a recent work, we used the DPCM method for coding the LARs, pitch and gain. Each of these transmission parameters was considered as a discrete-time signal with time instants given by the frame number. DPCM was applied to each of these signals independently of others.

We applied the DPCM method for coding the 14 transmission parameters (12 LARs, pitch and gain) extracted at a fixed rate of 50 frames/sec from 10 kHz sampled speech [6,14]. The resulting transmission bit rate was about 2000 bps. The DPCM coder of each parameter required the knowledge of its averaged standard deviation in order to compute the quantization step size employed by the coder. We observed improved speech quality either when this averaged standard deviation was updated by computing it over a current speech segment of about 2-3 seconds, or when an ADPCM coder (which adaptively changes the step size) was used.

With the use of our VFR transmission scheme, the correlation between adjacent transmitted frame data is greatly reduced, which means that the DPCM coder when used with the VFR scheme will yield little or no savings. Also, the above-mentioned 2000 bps DPCM-coded speech was found to have a slightly inferior overall quality compared to the speech at 1500 bps from an earlier version of our VFR system [1]. On the other hand, the DPCM coder has two main advantages: (1) it produces nearly fixed-rate bit stream, and (2) the hardware implementation of the DPCM coder and decoder is relatively simple and inexpensive.

10.2 Linear Predictive Formant Vocoder

It has been known for some time that formant vocoders enable speech transmission at very low bit rates (about 500 bps). One requires of these systems an acceptable level of speech intelligibility but not necessarily retention of naturalness of speaker characteristics. Such low-bit rate systems are of interest in some applications. Speech transmission through an underwater channel is a good example.

We conducted a preliminary experiment simulating a formant vocoder within our LPC system format. Formants were generated from LPC analysis data. The formant synthesizer was implemented not using resonators as in conventional formant vocoders, but employing the canonical or direct form realization of the linear

prediction all-pole filter. The predictor coefficients of the all-pole filter were computed from the received formant data. It is this difference in synthesizer implementation which enabled our formant vocoder to overcome some of the problems encountered by its predecessors. The LPC formant vocoder can accommodate variable number of formants in adjacent frames without causing any undesirable transients. Incorrect identification of formants, which in practice can occasionally happen due to imperfect formant tracking, produces less degradation in the quality of synthesized speech for the LPC formant vocoder than for its conventional counterparts. A third advantage stems from the result we reported in [1] that the parameters of the LPC synthesizer filter can be updated time-synchronously without introducing any transients. It is well-known that such transients occur if one updates the parameters of the resonators time-synchronously.

In the preliminary experiment, we employed the formant data already computed in our Speech Understanding Project. There, a 14-pole LPC analysis was done every 10 ms on speech sampled at 10 kHz and preemphasized using a 50 Hz first-order filter. The formant tracker used in that project then extracted, every 10 ms, up to a maximum of 3 formants in the frequency range 0-3100 Hz. For unvoiced sounds, often only two formants were determined. Gain and pitch were also computed every 10 ms. For the purposes

of the preliminary experiment, we did not quantize any of these analysis parameters. The receiver thus had a variable order LPC synthesizer. The synthesized speech was found to be quite intelligible except for the following type of problem: [s] was often perceived as [sh]. The reason for this problem is that [s] has significant energy concentration above 3.1 kHz unlike [sh] and that we essentially low-pass filtered speech at 3.1 kHz by considering only those formants below this frequency.

Encouraged by the results of the above work, we conducted a more detailed study of very low bit-rate speech compression systems, with support from ARPA-STO. The results of this work have been reported in [15].

REFERENCES

1. J. Makhoul, R. Viswanathan, L. Cosell and W. Russell, Natural Communication with Computers, Final Report, Vol. II, Speech Compression Research at BBN, Report No. 2976, Dec. 1974. (NTIS No. AD/A 003478/5GA, 104 pp.)

2. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-23, pp. 309-321, June 1975.

3. J. Makhoul and L. Cosell, "Recommendations for Encoding and Synthesis," ARPA NSC Note 49, Nov. 1974.

4. A.H. Gray, Jr. and J.D. Markel, "A Normalized Digital Filter Structure," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-23, pp. 268-277, June 1975.

5. J. Makhoul and L. Cosell, "Nothing to Lose, but Lot to Gain," ARPA NSC Note No. 56, March 1975.

6. BBN Quarterly Progress Report on Command and Control Related Computer Technology, BBN Report No. 3093, June 1975.

7. R. Viswanathan and J. Makhoul, "Specifications for ARPA-LPC System II," NSC Note 82, Feb. 1976.

8. R. Viswanathan, "Variable Frame Rate Transmission of Pitch and Gain," NSC Note 96, Sept. 1976.

9. L. Ehrman, "Analysis of some Redundancy Removal Bandwidth Compression Techniques," Proc. IEEE, Vol. 55, pp. 278-287, March 1967.

10. L.D. Davisson, "Data Compression Using Straight Line Interpolation," IEEE Trans. Inf. Theory, Vol. IT-14, pp. 390-394, May 1968.

11. R. Viswanathan, J. Makhoul and W. Russell, "Optimal Interpolation in Linear Predictive Vocoders," BBN Report No. 3065, April 1975 (also ARPA NSC Note No. 59).

12. L.R. Rabiner and R.E. Crochiere, "On the Design of All-Pass Signals with Peak Amplitude Constraints," Bell Syst. Tech. J., Vol. 55, pp. 395-407, April 1976.

13. BBN Quarterly Progress Report on Command and Control Related Computer Technology, BBN Report No. 3520, March 1977.

14. M.R. Sambur, "An Efficient Linear Prediction Vocoder," Bell Syst. Tech. J., Vol. 54, pp. 1693-1723, Dec. 1975.

15. J. Makhoul, C. Cook, R. Schwartz and D. Klatt, A Feasibility Study of Very Low Rate Speech Compression Systems, Final Report, BBN Report No. 3508, Feb. 1977.

16. J.D. Carroll, "Individual Differences and Multidimensional Scaling," in R.N. Shepard, A.K. Romney, and S. Nerlove (Eds) Multidimensional Scaling: Theory and applications in the behavioral sciences, Vol. 1, New York: Seminar Press, 1972, pp. 105-155.

17. M. Wish and J.D. Carroll "Applications of Individual Differences Scaling to Studies of Human Perception and Judgment," in E.C. Carterette and M.P. Friedman (Eds) Handbook of Perception, Volume 2: Psychophysical judgment and measurement. New York: Academic Press, 1974, pp. 449-491.

18. W.W. Cooley and P.R. Lohnes, Multivariate Procedures for the Behavioral Sciences, John Wiley and Sons, 1962, Chapter 3. (IBM Fortran Scientific Subroutine Package has a canonical correlation routine.)

19. BBN Quarterly Progress Report, Command and Control Related Computer Technology, BBN Report No. 3209, Dec. 1975.

20. K.N. Stevens, M.H.L. Hecker and K.D. Kryter, "An Evaluation of Speech Compression Systems," BBN Report No. 914, March 1962.

21. K.N. Stevens, "Simplified Nonsense-Syllable Tests for Analytic Evaluation of Speech Transmission Systems," J. Acoust. Soc. Amer., Vol. 34, p. 729, May 1962.

22. W.D. Voiers, D. Sharpley and C.J. Hehmsoth, "Research on Diagnostic Evaluation of Speech Intelligibility," Report No. AFCRL-72-0694, Sept. 1972.

23. BBN Quarterly Progress Report, Command and Control Related Computer Technology, BBN Report No. 3263, March 1976.

24. P.T. Brady, "Effects of Transmission Delay on Conversational Behavior on Echo-free Telephone Circuits," Bell Syst. Tech. J., Vol. 50, pp. 115-134, 1971.

25. A.W.F. Huggins, "Effect of Lost Packets on Speech Intelligibility," NSC Note 78, Feb. 1976.

26. A.W.F. Huggins, "Temporally Segmented Speech," Perception and Psychophysics, Vol. 18, pp. 149-157, 1975.

27. W.D. Voiers, "Methods of Predicting User Acceptance of Voice Communication Systems," Final Report for DCA, July 1976.

28. R.G. Crowder and J. Morton, "Precategorical Acoustic Storage (PAS)," Perception and Psychophysics, Vol. 5, pp. 365-373, 1969.

29. R.G. Crowder, "Audition And Speech Coding in Short-Term Memory: a Tutorial Review," Presented at Attention and Performance VII, Senanque, France, August 1976.

30. D.N. Kalikow, K.N. Stevens, and L.L. Elliott, "Development of a Test of Speech Intelligibility in Noise Using Sentence Materials with Controlled Word Predictability," J. Acoust. Soc. Amer., Vol. 61, pp. 1337-1351, May 1977.

APPENDIX 1

STABLE AND EFFICIENT LATTICE METHODS FOR LINEAR PREDICTION

(Paper published in IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-25, Oct. 1977.)

Stable and Efficient Lattice Methods for Linear Prediction

JOHN MAKHOUL, MEMBER, IEEE

Abstract—A class of stable and efficient recursive lattice methods for linear prediction is presented. These methods guarantee the stability of the all-pole filter, with or without windowing of the signal, with finite wordlength computations, and at a computational cost comparable to the traditional autocorrelation and covariance methods. In addition, for data-compression purposes, quantization of the reflection coefficients can be accomplished within the recursion, if desired.

I. Introduction

THE autocorrelation method of linear prediction [1] guarantees the stability of the all-pole filter, but has the disadvantage that windowing of the signal causes a reduction in spectral resolution. In practice, even the stability is not always guaranteed with finite wordlength (FWL) computations [2]. On the other hand, the covariance method [1], [3] does not guarantee the stability of the filter, even with floating-point computation, but has the advantage that there is no windowing of the signal. One solution to these problems was given by Itakura [4] in his lattice formulation. In this method, filter stability is guaranteed with no windowing and with much smaller sensitivity to FWL computations. Unfortunately, this is accomplished with about a fourfold increase in computation over the other two methods. A similar method was independently proposed by Burg [5], [6].

This paper presents a class of lattice methods that guarantees the stability of the all-pole filter, independently of the stationarity properties and the duration of the signal. It is shown that the methods of Itakura and Burg are special cases of this class of methods. Furthermore, a procedure is given that reduces the number of computations to values comparable to those in the autocorrelation and covariance methods. In this procedure, the "forward" and "backward" residuals are not computed; the reflection coefficients are computed directly from the covariance of the input signal.

Section II presents the class of lattice methods for computing the reflection coefficients, along with conditions for ensuring stability. Section III describes a procedure, termed the *covariance-lattice* method, for performing the necessary computations efficiently. Computational issues are then discussed in Section IV, followed in Section V by a step-by-step procedure for one of the promising lattice methods for linear predictive analysis.

Manuscript received May 5, 1976; revised September 9, 1976, and May 13, 1977. This work was supported by the Information Processing Techniques Branch of the Advanced Research Projects Agency under Contracts MDA903-75-C-0180 and N00014-75-C-0533.

The author is with Bolt Beranek and Newman Inc., Cambridge, MA 02138.

II. Lattice Formulations

In linear prediction, the signal spectrum is modeled by an all-pole spectrum with a transfer function given by

$$H(z) = \frac{G}{A(z)} \tag{1}$$

where

$$A(z) = \sum_{k=0}^{p} a_k z^{-k}, \quad a_0 = 1 \tag{2}$$

is known as the inverse filter, G is a gain factor, a_k are the predictor coefficients, and p is the number of poles or predictor coefficients in the model. If $H(z)$ is stable (minimum phase), $A(z)$ can be implemented as a lattice filter [4], as shown in Fig. 1. The reflection (or partial correlation) coefficients K_m in the lattice are uniquely related to the predictor coefficients. Given K_m, $1 \leq m \leq p$, the set $\{a_k\}$ is computed by the recursive relation

$$\begin{aligned} a_m^{(m)} &= K_m \\ a_j^{(m)} &= a_j^{(m-1)} + K_m a_{m-j}^{(m-1)}, \quad 1 \leq j \leq m-1, \end{aligned} \tag{3}$$

where the equations in (3) are computed recursively for $m = 1, 2, \cdots, p$. After each recursion, the coefficients $a_j^{(m)}$, $1 \leq j \leq m$, are the desired coefficients for the mth-order predictor. The final solution is given by $a_j = a_j^{(p)}$, $1 \leq j \leq p$. For a stable $H(z)$, one must have

$$|K_m| < 1, \quad 1 \leq m \leq p. \tag{4}$$

In the lattice formulation, the reflection coefficients can be computed by minimizing some norm of the forward residual $f_m(n)$ or the backward residual $b_m(n)$, or a combination of the two. From Fig. 1, the following relations hold:

$$f_0(n) = b_0(n) = s(n) \tag{5a}$$

$$f_{m+1}(n) = f_m(n) + K_{m+1} b_m(n-1) \tag{5b}$$

$$b_{m+1}(n) = K_{m+1} f_m(n) + b_m(n-1) \tag{5c}$$

where $s(n)$ is the input signal and $e(n) = f_p(n)$ is the output residual. In z-transform notation: $E(z) = A(z)S(z)$.

We shall give several methods for the determination of the reflection coefficients. These methods depend on different ways of correlating the forward and backward residuals. Below, we shall make use of the following definitions:

$$F_m(n) = E[f_m^2(n)] \tag{6a}$$

$$B_m(n) = E[b_m^2(n)] \tag{6b}$$

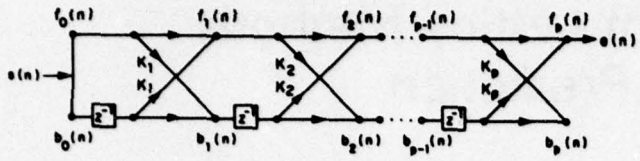


Fig. 1. Lattice inverse filter $A(z)$.

$$C_m(n) = E[f_m(n)b_m(n-1)], \tag{6c}$$

where $E(\cdot)$ denotes the expected value. The left-hand side of each of the equations in (6) is a function of n because we are making the general assumption that the signals are nonstationary. (Subscripts, etc., will be dropped sometimes for convenience.)

A. Forward Method

In this method, the reflection coefficient at stage $m+1$ is obtained as a result of the minimization of an error norm given by the variance (or mean square) of the forward residual

$$F_{m+1}(n) = E[f_{m+1}^2(n)]. \tag{7}$$

By substituting (5b) in (7) and differentiating with respect to K_{m+1}, one obtains

$$K_{m+1}^f = -\frac{E[f_m(n)b_m(n-1)]}{E[b_m^2(n-1)]} = -\frac{C_m(n)}{B_m(n-1)}. \tag{8}$$

This method of computing the filter parameters is similar to the autocorrelation and covariance methods in that the mean-squared *forward* residual is minimized.

B. Backward Method

In this case, the minimization is performed on the variance of the backward residual at stage $m+1$. From (5c) and (6b), the minimization of $B_{m+1}(n)$ leads to

$$K_{m+1}^b = -\frac{E[f_m(n)b_m(n-1)]}{E[f_m^2(n)]} = -\frac{C_m(n)}{F_m(n)}. \tag{9}$$

Note that, since $F_m(n)$ and $B_m(n-1)$ are both nonnegative and the numerators in (8) and (9) are identical, K^f and K^b always have the same sign S

$$S = \text{sign } K^f = \text{sign } K^b. \tag{10}$$

C. Geometric-Mean Method (Itakura)

The main problem in the previous two techniques is that the computed reflection coefficients are not always guaranteed to be less than 1 in magnitude; i.e., the stability of $H(z)$ is not guaranteed. One solution to this problem was offered by Itakura [4] where the reflection coefficients are computed from

$$K_{m+1}^I = -\frac{E[f_m(n)b_m(n-1)]}{\sqrt{E[f_m^2(n)]\,E[b_m^2(n-1)]}} = -\frac{C_m(n)}{\sqrt{F_m(n)B_m(n-1)}}. \tag{11}$$

K_{m+1}^I is the negative of the statistical correlation between $f_m(n)$ and $b_m(n-1)$; hence, property (4) follows. To the author's knowledge, (11) cannot be derived directly by minimizing some error criterion. However, from (8), (9), and (11), one can easily show that K^I is the geometric mean of K^f and K^b

$$K^I = S\sqrt{K^f K^b} \tag{12}$$

where S is given by (10), and we have omitted the subscript $m+1$. From the properties of the geometric mean, it follows that

$$\min[|K^f|, |K^b|] \leq |K^I| \leq \max[|K^f|, |K^b|].$$

Now, since $|K^I| < 1$, it follows that *if the magnitude of either K^f or K^b is greater than 1, the magnitude of the other is necessarily less than 1.* This important property can be summarized by the following.

$$\text{If } |K^f| > 1, \text{ then } |K^b| < 1,$$

or

$$\text{if } |K^b| > 1, \text{ then } |K^f| < 1. \tag{13}$$

Property (13) immediately brings to mind another possible definition for the reflection coefficient that guarantees stability.

D. Minimum Method

$$K^M = S \min[|K^f|, |K^b|]. \tag{14}$$

This says that at each stage, compute K^f and K^b and choose as the reflection coefficient the one with the smaller magnitude. Property (13) guarantees that K^M satisfies (4).

E. General Method

Between K^M and K^I there are an infinity of values that can be chosen as valid reflection coefficients (i.e., $|K| < 1$). These can be conveniently defined by taking the generalized rth mean of K^f and K^b

$$K^r = S\left[\tfrac{1}{2}(|K^f|^r + |K^b|^r)\right]^{1/r}. \tag{15}$$

As $r \to 0$, $K^r \to K^I$, the geometric mean. For $r > 0$, K^r cannot be guaranteed to satisfy (4). Therefore, for K^r to be a reflection coefficient, we must have $r \leq 0$. In particular

$$K^0 = K^I, \quad K^{-\infty} = K^M. \tag{16}$$

If the signal is stationary, one can show that $K^f = K^b$, and that

$$K^r = K^f = K^b, \quad \text{all } r \text{ (Stationary Case)}. \tag{17}$$

F. Harmonic-Mean Method (Burg)

There is one value of r for which K^r has some interesting properties, and that is $r = -1$. K^{-1}, then, would be the harmonic mean of K^f and K^b

$$K_{m+1}^B = K^{-1} = \frac{2K^f K^b}{K^f + K^b} = -\frac{2C_m(n)}{F_m(n) + B_m(n-1)}. \tag{18}$$

One can show that

$$|K^M| \leq |K^B| \leq |K^I|. \tag{19}$$

TABLE I
COMPUTATIONAL COST FOR TRADITIONAL LINEAR-PREDICTION METHODS AS COMPARED TO THE NEW AUTOCORRELATION-LATTICE AND COVARIANCE-LATTICE METHODS

	AUTOCORRELATION METHOD	COVARIANCE METHOD	REGULAR LATTICE (WITH RESIDUALS)
TRADITIONAL METHODS	$pN + p^2$	$pN + \frac{1}{6}p^3 + \frac{3}{2}p^2$	$5pN$
NEW LATTICE METHODS	$pN + \frac{1}{6}p^3 + \frac{3}{2}p^2$	$pN + \frac{1}{2}p^3 + 2p^2$	$5pN$

Note: Terms of order p have been neglected.

In fact, Itakura used K^B as an approximation to K^I in (11) to avoid computing the square root.

One important property of K^B that is not shared by K^I and K^M, is that K^B results directly from the minimization of an error criterion. The error is defined as the sum of the variances of the forward and backward residuals

$$E_{m+1}(n) = F_{m+1}(n) + B_{m+1}(n). \tag{20}$$

Using (5) and (6), one can show that the minimization of (20) indeed leads to (18). One can also show that the forward and backward minimum errors at stage $m + 1$ are related to those at stage m by the following:

$$F_{m+1}(n) = [1 - (K^B_{m+1})^2]\, F_m(n) \tag{21a}$$

$$B_{m+1}(n) = [1 - (K^B_{m+1})^2]\, B_m(n-1). \tag{21b}$$

This formulation is originally due to Burg [5], [6].

G. Discussion

Note that, in general, lattice methods do not minimize any global error criterion, such as the variance of the final forward residual, etc. Any minimization that might take place is done stage by stage. If the signal $s(n)$ is truly stationary, the stage-by-stage minimization gives the same result as global minimization. In fact, for a stationary signal, all the lattice methods previously described, as well as the autocorrelation and covariance methods, give the same result. However, in general, the signal cannot be assumed to be stationary and the different lattice methods will give different results, which are still different from the covariance-method result. The lattice methods will indeed give suboptimal solutions; solutions that tend to an optimal solution as the signal becomes more stationary. Which lattice method to choose in a particular situation, then, is not clear cut. We tend to prefer the use of K^B in (18) because it minimizes a reasonable and well-defined error criterion.

III. THE COVARIANCE-LATTICE METHOD

If linear predictive analysis is to be performed on a regular computer, the number of computations for the lattice methods given far exceeds that of the autocorrelation and covariance methods (see the first row of Table I). This is unfortunate since, otherwise, lattice methods generally have superior properties when compared to the autocorrelation and covariance methods (see Table II). Below, we derive a new method, called the *covariance-lattice* method, which has all the advantages of a regular lattice, but with an efficiency comparable to the two nonlattice methods.

TABLE II
COMPARISON BETWEEN DIFFERENT PROPERTIES OF VARIOUS LINEAR-PREDICTION METHODS

PROPERTY	LINEAR PREDICTION METHOD			
	AUTOCORRELATION	COVARIANCE	REGULAR LATTICE	COVARIANCE LATTICE
WINDOWING	NECESSARY	NONE	NOT NECESSARY	NONE
STABILITY	THEORETICALLY GUARANTEED	NOT GUARANTEED	CAN BE GUARANTEED	
STABILITY WITH FINITE WORDLENGTH COMPUTATIONS	NOT GUARANTEED		CAN BE GUARANTEED	
COMPUTATIONAL EFFICIENCY	EFFICIENT		EXPENSIVE	EFFICIENT
LEAST-SQUARES OPTIMALITY	OPTIMAL		POSSIBLY ONLY SUBOPTIMAL	
QUANTIZATION OF REFLECTION COEFFICIENTS WITHIN RECURSION	NOT POSSIBLE		POSSIBLE	
NUMBER OF SAMPLES FOR ANALYSIS	N	CAN BE REDUCED TO ~0.7N FOR THE SAME RESOLUTION		

From the recursive relations in (3) and (5), one can show that

$$f_m(n) = \sum_{k=0}^{m} a_k^{(m)} s(n-k) \tag{22a}$$

$$b_m(n) = \sum_{k=0}^{m} a_k^{(m)} s(n-m+k). \tag{22b}$$

Squaring (22a) and taking the expected value, there results

$$F_m(n) = \sum_{k=0}^{m} \sum_{i=0}^{m} a_k^{(m)} a_i^{(m)} \phi(k,i) \tag{23}$$

where

$$\phi(k,i) = E[s(n-k)\, s(n-i)] \tag{24}$$

is the nonstationary autocorrelation (or covariance) of the signal $s(n)$. ($\phi(k,i)$ in (24) is technically a function of n, which has been dropped for convenience.) In a similar fashion one can show from (22b), with n replaced by $n-1$, that

$$B_m(n-1) = \sum_{k=0}^{m} \sum_{i=0}^{m} a_k^{(m)} a_i^{(m)} \phi(m+1-k, m+1-i) \tag{25}$$

$$C_m(n) = \sum_{k=0}^{m} \sum_{i=0}^{m} a_k^{(m)} a_i^{(m)} \phi(k, m+1-i). \tag{26}$$

Given the covariance of the signal, the reflection coefficient at stage $m + 1$ can be computed from (23), (25), and (26) by substituting them in the desired formula for K_{m+1}. The name "covariance-lattice" stems from the fact that this is basically a lattice method that is computed from the covariance of the signal; it can be viewed as a way of stabilizing the covariance method. One salient feature is that the forward and backward residuals are never actually computed in this method. But this is not different from the nonlattice methods.

In the harmonic-mean method (18), $F_m(n)$ need not be computed from (23); one can use (21a) instead, with m replaced by $m-1$. However, one must use (25) to compute $B_m(n-1)$;

(21b) cannot be used because $B_{m-1}(n-2)$ would be needed and it is not readily available.

A. Stationary Case

For a stationary signal, the covariance reduces to the autocorrelation

$$\phi(k,i) = R(i-k) = R(k-i) \qquad \text{(Stationary)}. \quad (27)$$

From (23)-(27), it is clear that

$$F_m = B_m = \sum_{k=0}^{m} \sum_{i=0}^{m} a_k^{(m)} a_i^{(m)} R(i-k) \quad (28)$$

and

$$C_m = \sum_{k=0}^{m} \sum_{i=0}^{m} a_k^{(m)} a_i^{(m)} R(m+1-i-k). \quad (29)$$

Making use of the normal equations [1]

$$\sum_{i=0}^{m} a_i^{(m)} R(i-k) = 0, \quad 1 \leq k \leq m \quad (30)$$

and of (21), one can show that the stationary reflection coefficient is given by

$$K_{m+1} = -\frac{C_m}{F_m} = -\frac{\sum_{k=0}^{m} a_k^{(m)} R(m+1-k)}{(1-K_m^2)F_{m-1}} \quad (31)$$

with $F_0 = R_0$. Equation (31) is exactly the equation used in the autocorrelation method.

B. Quantization of Reflection Coefficients

One of the features of lattice methods is that the quantization of the reflection coefficients can be accomplished within the recursion, i.e., K_m can be quantized before K_{m+1} is computed. In this manner, it is hoped that some of the effects of quantization can be compensated for.

In applying the covariance-lattice procedure to the harmonic-mean method, one must be careful to use (23) and *not* (21a) to compute $F_m(n)$. The reason is that (21a) is based on the optimality of K^B, which would no longer be true after quantization.

Similar reasoning can be applied to the autocorrelation method. Those who have tried to quantize K_m inside the recursion have no doubt been met with serious difficulties. The reason is that (31) assumes the optimality of the predictor coefficients at stage m, which no longer would be true if K_m were quantized. The solution is to use (28) and (29), which make no assumptions of optimality. Thus we have what we shall call the *autocorrelation-lattice* method, where there is only one definition of K_{m+1}

$$K_{m+1} = -\frac{C_m}{F_m} = -\frac{\sum_{k=0}^{m} \sum_{i=0}^{m} a_k^{(m)} a_i^{(m)} R(m+1-i-k)}{\sum_{k=0}^{m} \sum_{i=0}^{m} a_k^{(m)} a_i^{(m)} R(i-k)}. \quad (32)$$

IV. Computational Issues

A. Simplifications

Equations (23), (25), and (26) can be rewritten to reduce the number of computations by about one half. The results for $C_m(n)$ and $F_m(n) + B_m(n-1)$ can be shown to be as follows:

$$C_m(n) = \phi(0, m+1) + \sum_{k=1}^{m} a_k^{(m)} [\phi(0, m+1-k) + \phi(k, m+1)] + \sum_{k=1}^{m} [a_k^{(m)}]^2 \phi(k, m+1-k) + \sum_{k=1}^{m-1} \sum_{i=k+1}^{m} a_k^{(m)} a_i^{(m)} [\phi(k, m+1-i) + \phi(i, m+1-k)] \quad (33)$$

$$F_m(n) + B_m(n-1) = \phi(0,0) + \phi(m+1, m+1) + 2\sum_{k=1}^{m} a_k^{(m)} [\phi(0,k) + \phi(m+1, m+1-k)] + \sum_{k=1}^{m} [a_k^{(m)}]^2 [\phi(k,k) + \phi(m+1-k, m+1-k)] + 2\sum_{k=1}^{m-1} \sum_{i=k+1}^{m} a_k^{(m)} a_i^{(m)} \cdot [\phi(k,i) + \phi(m+1-k, m+1-i)]. \quad (34)$$

The third term in (33) can be computed more efficiently as follows:

$$\sum_{k=1}^{m} [a_k^{(m)}]^2 \phi(k, m+1-k) = \sum_{k=1}^{m/2} \{[a_k^{(m)}]^2 + [a_{m+1-k}^{(m)}]^2\} \phi(k, m+1-k) + \underbrace{[a_{(m+1)/2}^{(m)}]^2 \phi\left(\frac{m+1}{2}, \frac{m+1}{2}\right)}_{\text{only if } m \text{ odd}}. \quad (35)$$

A similar simplification can be used in (34).

For the stationary case, (28) can be rewritten as

$$F_m = \sum_{k=-m}^{m} b_k R(k) = b_0 + 2\sum_{k=1}^{m} b_k R(k) \quad (36)$$

where

$$b_k = \sum_{l=0}^{m-|k|} a_l^{(m)} a_{l+|k|}^{(m)} \quad (37)$$

is the autocorrelation of the impulse response of $A(z)$. By setting $l = i + k$ in (29), one can show that C_m is reduced to

$$C_m = \sum_{l=0}^{2m} c_l R(m+1-l) \quad (38)$$

where

$$c_l = \sum_{k=0}^{m} a_k^{(m)} a_{l-k}^{(m)}, \quad 0 \leq l \leq 2m \tag{39}$$

is the convolution of the impulse response of $A(z)$ with itself. Equation (39) assumes that $a_k^{(m)} = 0$ for $k < 0$ and $k > m$. Equation (38) can be rewritten as

$$C_m = R(m+1) + 2a_1^{(m)} R(m) + c_{m+1} R(0) + \sum_{k=1}^{m-1} (c_{m+1-k} + c_{m+1+k}) R(k). \tag{40}$$

Equation (39) can also be rewritten to reduce the computations further.

B. Covariance Computation

The covariance $\phi(k, i)$ of the signal is defined in (24) as a nonstationary autocorrelation, which, strictly speaking, should be estimated by averaging over an ensemble of the random process. In practice, however, it is often the case that such averaging is neither feasible nor desirable. For example, in most speech applications, one is interested in analyzing the time-varying properties of a particular utterance and not the whole ensemble of speech that a speaker might utter. In the case where a single time history of a random process is available for analysis, it is common to describe that single time record as nonstationary if its short-term sample properties (such as mean and autocorrelation) vary significantly with time [8]. For this situation, we give below two methods for computing the covariance of a signal that is known, say, for $0 \leq n \leq N-1$.

Method 1:

$$\phi(k, i) = \sum_{n=p}^{N-1} s(n-k)s(n-i), \quad 0 \leq k, i \leq p \tag{41}$$

where p is the order of the predictor, and the customary division by the number of terms in the summation (in this case $N-p$) has been omitted since it does not affect the solution for the reflection coefficients. If we assume that (41) estimates the covariance at time $t = 0$, then the covariance at any other time t can be estimated by setting the lower and upper limits of the summation in (41) to $p + t$ and $N - 1 + t$, respectively. Note that (41) makes no assumptions about the signal outside the given range and, hence, is especially useful for short durations [6] and nonstationary signals. On the other hand, if the signal is assumed to be zero outside the given range (i.e., the signal is windowed), then the signal is effectively forced to be stationary, with an associated autocorrelation given by

$$R(i) = \sum_{n=0}^{N-1-|i|} s_n s_{n+|i|}, \quad 0 \leq i \leq p. \tag{42}$$

Method 2: The second method makes maximum use of the data in the range $0 \leq n \leq N-1$. This is accomplished by recomputing the covariance for each new lattice stage as follows:

$$\phi_m(k, i) = \sum_{n=m}^{N-1} s(n-k)s(n-i), \quad 0 \leq k, i \leq m \tag{43}$$

where $\phi_m(k, i)$ is the covariance used in computing K_m. The computations in (43) can be simplified considerably by noting that

$$\phi_{m+1}(k, i) = \phi_m(k, i) - s(m-k)s(m-i), \quad 0 \leq k, i \leq m. \tag{44}$$

Therefore, the covariance coefficients for stage $m+1$ can be computed from those for stage m using (44) in the range $0 \leq k, i \leq m$. For $k = m+1$ or $i = m+1$, (43) needs to be used.

It can be shown that when Method 2 for computing the covariance is used in conjunction with the harmonic-mean computation in (18), the results for the reflection coefficients are identical to Burg's method as described in [6]. However, our results here are obtained at a much lower computational cost.

For the case where $N >> p$, Methods 1 and 2 should give similar results. However, if N is not much greater than p, then it would seem reasonable to utilize the given data maximally by using Method 2.

There are other possible methods for computing the covariance or the autocorrelation of the signal. Irrespective of which method one chooses, it is important to make sure that the resulting covariance or autocorrelation function is positive definite. Otherwise, filter stability cannot be guaranteed.

C. Computational Cost

Table I shows a comparison of the number of computations for the different methods, where terms of order p have been neglected. The computations for the autocorrelation-lattice and covariance-lattice methods are on the order of $pN + O(p^3)$, as compared to $5pN$ for the regular lattice methods where the residuals are computed. For $N >> p$, the new lattice methods typically offer a 3–4-fold saving over the regular lattice methods.

When compared to nonlattice methods, the increase in computation for the covariance-lattice method is not significant if N is large compared to p, which is usually the case (compare the first and second rows in Table I). Furthermore, in the covariance-lattice method, the number of signal samples can be reduced to about half that used in the autocorrelation method. This not only reduces the number of computations but also improves spectral resolution by reducing the amount of averaging.

D. FWL Computations

One point of comparison between the different methods is the stability of the all-pole filter when FWL computations are used. The main comparison here is between the autocorrelation method and the lattice methods (the covariance method cannot guarantee stability, in general, even with floating-point computations). Under FWL conditions, we expect filter stability to be ensured more with the lattice methods than with the autocorrelation method. If, at some stage of the recursion, K_m turns out to be greater than one because of FWL computations, it can be artificially set to a value less than one to ensure stability. Such a scheme would work well with the lattice methods, but not with the autocorrelation method because in the latter, global optimality of each K is assumed at every

stage. Lack of optimality leads to error propagation, which in turn makes later stages more susceptible to instability. The problem does not exist to the same magnitude in the lattice methods since consecutive stages are "decoupled," with no assumptions of global optimality being made. This phenomenon is the same as that discussed in Section III-B, which allows the quantization of the reflection coefficient inside the recursion of the lattice methods.

V. Procedure

Below is the complete algorithm for what we believe currently to be one of the more promising methods for linear predictive analysis. It comprises the harmonic-mean definition (18) for the reflection coefficients, and the covariance-lattice method.

a) Compute the covariances $\phi(k, i)$ for $k, i = 0, 1, \cdots, p$.
b) $m \leftarrow 0$.
c) Compute $C_m(n)$ and $F_m(n) + B_m(n-1)$ from (33) and (34), or from (23), (25), and (26).
d) Compute K_{m+1} from (18).
e) Quantize K_{m+1}, if desired (perhaps using log area ratios [7] or some other technique).
f) Using (3), compute the predictor coefficients $\{a_k^{(m+1)}\}$ from $\{a_k^{(m)}\}$ and K_{m+1}. Use the quantized value, if K_{m+1} was quantized in d).
g) $m \leftarrow m + 1$.
h) If $m < p$, go to c); otherwise exit.

We have used this procedure to analyze speech signals, with the signal covariance estimated by (41). In general, the results were somewhere between those using the autocorrelation and covariance methods. In particular, the pole bandwidths were usually less than those from the autocorrelation method, but greater than those from the covariance method. In all cases where the covariance method gave unstable results, the covariance-lattice method gave stable results.

While the performance of all linear prediction methods tends to deteriorate (in terms of spectral accuracy) as the number of signal samples N is sharply reduced, we believe that the procedure given above should continue to give better resolution than the autocorrelation method, and should continue to guarantee stability, unlike the covariance method, which tends to become unstable for short durations.

VI. Conclusions

This paper presented a class of lattice methods for linear prediction that guarantees the stability of the all-pole filter, with or without windowing of the signal, and with FWL computations. Also, for data-compression purposes, quantization of the reflection coefficients can be accomplished within the recursion, if desired, without affecting the stability of the filter. It was shown that the methods of Itakura and Burg are special cases of this class of lattice methods.

A procedure was derived to make these lattice methods more efficient computationally, with a cost comparable to the traditional autocorrelation and covariance methods. The procedure, named the *covariance-lattice* method, computes the reflection coefficients recursively in terms of the covariance of the signal and the filter parameters at each stage. When used with speech signals, this method gave results somewhere in between the autocorrelation and covariance methods.

Acknowledgment

The author wishes to thank R. Viswanathan for implementing the covariance-lattice method and for his discussions and comments on this paper.

References

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[2] J. Markel and A. H. Gray, Jr., "Fixed-point truncation arithmetic implementation of a linear prediction autocorrelation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 273–281, Apr. 1974.
[3] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, Aug. 1971.
[4] F. Itakura and S. Saito, "Digital filtering techniques for speech analysis and synthesis," presented at the 7th Int. Cong. Acoustics, Budapest, 1971, Paper 25-C-1.
[5] J. Burg, "Maximum entropy spectral analysis," Ph. D. dissertation, Stanford Univ., Stanford, CA, May 1975.
[6] D. E. Smylie, G. K. C. Clarke, and T. J. Ulrych, "Analysis of Irregularities in the earth's rotation," in *Methods in Computational Physics*, vol. 13. New York: Academic, 1973, pp. 391–430.
[7] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 309–321, June 1975.
[8] J. Bendat and A. Piersol, *Random Data: Analysis and Measurement Procedures*. New York: Wiley-Interscience, 1971, pp. 13–14.

APPENDIX 2

SEQUENTIAL LATTICE METHODS FOR STABLE LINEAR PREDICTION

(Paper presented at the 1976 EASCON Conference, Washington, D.C., Sept. 1976.)

SEQUENTIAL LATTICE METHODS FOR STABLE LINEAR PREDICTION

R. VISWANATHAN and J. MAKHOUL

Bolt Beranek and Newman Inc., Cambridge, Mass. 02138

ABSTRACT

A sequential linear prediction method computes new values for the parameters of the predictor on a sample-by-sample basis. Under the assumption of an all-pole (or autoregressive) model, a number of methods are developed in this paper for sequentially estimating the model parameters. A common thread in all the developed methods is that they employ the lattice model of the linear prediction filter and that they all guarantee the filter stability. Several applications of sequential estimation are considered in speech signal processing. While the paper contains mainly theoretical developments, results of experimental investigations of the reported methods will be presented at the conference.

1. INTRODUCTION

Recently a class of lattice methods were proposed for linear prediction with the resulting all-pole filter guaranteed to be stable [1]. In this paper, we extend these methods to permit sequential estimation. A sequential method, by our definition, provides a new estimate for the filter coefficients upon receiving each signal sample. Below we limit our discussion to the all-pole (or autoregressive) model, and consider applications in speech signal processing.

Before we consider sequential linear prediction methods, we review the lattice formulation for block linear prediction in Section 2. (A block or batch-processing method provides one estimate for the filter coefficients over a given block of signal samples.) The types of sequential methods developed in this paper are described in Section 3. From an operational viewpoint, these methods are grouped into two classes: (1) Block sequential estimation, and (2) Recursive estimation. Based on the time extent of dependence of the present estimate on past signal samples, sequential methods are grouped into three classes: (a) Fixed (finite) memory, (b) Growing memory, and (c) Fading memory. Section 4 deals with block sequential estimation and Section 5, with recursive estimation. Both sections treat the three memory conditions given above. Section 6 points out two important differences between block sequential and recursive estimation approaches.

Sequential methods require, in general, increased computation compared to block methods. There are, however, a number of potential advantages in having the filter coefficients available on a sample-by-sample basis [2-4,14-17]. These advantages, as applied to speech signal processing, are considered in Section 7.

2. LATTICE FORMULATION FOR LINEAR PREDICTION

The lattice formulation was introduced in speech by Itakura [7], and in geophysics, by Burg [8]. (Burg's method is known as the maximum entropy method.) Recently, Makhoul showed the existence of a class of such lattice methods all of which guarantee the stability of the all-pole filter, with or without windowing of the signal; also, stability is less sensitive to finite wordlength computations [1]. Unfortunately, these methods (hereafter called regular lattice methods) cause about a four-fold increase in computation over the traditional autocorrelation and covariance methods [9]. To overcome this drawback, Makhoul introduced the so-called covariance lattice methods: these compute the lattice model parameters directly from the covariance of the signal, and thus require about the same order of computational complexity as the two traditional methods [1]. Since our purpose is to extend both the regular and covariance lattice methods to permit sequential linear prediction , we shall next explain the lattice model and introduce the necessary terminology.

In linear prediction, the signal spectrum is modelled by an all-pole spectrum with a transfer function given by

$$H(z) = G/A(z), \tag{1}$$

where $$A(z) = \sum_{k=0}^{p} a_k z^{-k}, \quad a_0 = 1, \tag{2}$$

is known as the inverse filter, G is a gain factor, a_k are the predictor coefficients, and p is the number of poles or predictor coefficients in the model. If H(z) is stable, A(z) can be implemented as a lattice filter, as shown in Fig. 1. The reflection (or partial correlation) coefficients K_m in the lattice are uniquely related to the predictor coefficients. For a stable H(z), one must have

$$|K_m| < 1, \quad 1 \leq m \leq p. \tag{3}$$

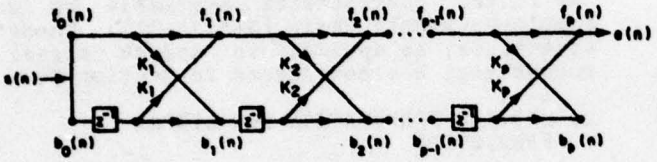


Fig. 1. Lattice inverse filter.

In the lattice formulation, the reflection coefficients can be computed by minimizing some norm of the forward residual $f_m(n)$ or the backward residual $b_m(n)$, or a combination of the two. From Fig. 1, the following relations hold:

$$f_0(n) = b_0(n) = s(n), \tag{4a}$$

$$f_{m+1}(n) = f_m(n) + K_{m+1}\, b_m(n-1), \tag{4b}$$

$$b_{m+1}(n) = K_{m+1}\, f_m(n) + b_m(n-1). \tag{4c}$$

s(n) is the input signal and $e(n)=f_p(n)$ is the output residual.

There are a number of methods for estimating the reflection coefficients which satisfy the stability condition (3). Each of these methods may be extended to perform sequential estimation.

Besides the important stability consideration, there are other factors that favor employing the lattice model in general. Lattice form implementation of (1) produces a lower sensitivity to roundoff noise than, for example, the direct form implementation [10]. The reflection coefficients, which are the parameters of the lattice model, were found to be the best for use in speech transmission systems [11]. Also, the reflection coefficients have an orthogonality property in the sense that an (m+1)-stage lattice has its first m reflection coefficients identical to those of the m-stage lattice. Using this property and a suitable criterion, an estimate of the "true" order of the model for a given signal sequence may be readily obtained [9,12]. In fact, such an estimate was employed in the design of variable order linear prediction as a data compression technique [6].

3. TYPES OF SEQUENTIAL ESTIMATION METHODS

Sequential estimation methods presented in this paper can be classified in two different ways, first by considering the operational aspect of the estimator, and second based on estimator memory.

From an operational viewpoint, we have two classes of sequential methods:

(1) Block sequential estimation,
(2) Recursive estimation.

A block sequential estimator provides sample-by-sample estimates by successively *applying a block linear prediction method.* Since, for block linear prediction, covariance lattice methods give the same results as regular lattice methods, but at substantial computational savings, we exclusively consider the use of covariance lattice methods in block sequential estimation. A recursive estimator determines a new estimate at time n as a function of the last estimate at time n-1 and a quantity that is available at time n. (This latter quantity may be called a "measurement" at time n, following the control theory or Kalman filter terminology.) Regular lattice methods and a version of Widrow's least mean squares method are considered as examples of recursive estimation.

Based on the nature of the estimator memory, we group sequential methods into three classes. By memory, we mean the dependence (direct or indirect) of the current estimate on past signal samples. The three classes are:

(a) Fixed memory methods,
(b) Growing memory methods,
(c) Fading memory methods.

The extent of the estimator memory is constrained to be constant for class (a); as new signal samples arrive, the estimator memory is updated such that the signal samples furthest in the past are

discarded to make room for the most recent signal samples. For class (b), the size of the estimator memory increases as new data is processed. Fading memory methods, which form class (c), can have either a fixed or growing memory span, but the most recent data is given greater emphasis than the data further back in time.

Section 4 below deals with block sequential estimation, and Section 5, with recursive estimation. Both sections treat methods from each of the three memory-based classes given above.

4. BLOCK SEQUENTIAL ESTIMATION (BSE)

Upon receiving a sample s(n), a BSE method finds a stable estimate $\{K_m(n)\}$ in two steps as follows. First, from its memory span {s(n), s(n-1),...} (which has a constant or increasing number of samples depending upon whether the BSE method has a fixed or growing memory), it computes the covariance matrix

$$\Phi(n) = [\phi(i,j,n)], \quad 0 \le i,j \le p, \tag{5}$$

where $\phi(i,j,n)$ is the i-jth covariance at time n. The second step is to apply any of the covariance lattice methods given in [1] to solve for the lattice parameters $K_m(n)$. We show next that, under each of the three memory conditions, computing $\Phi(n)$ can be accomplished at significant computational savings by making use of the knowledge of $\Phi(n-1)$. (Of course, at the very beginning of the signal sequence where the estimator is just starting up, the covariance matrix has to be computed directly from signal samples. In fact, in that initial period, the first estimate is available only after a certain number of signal samples have been accumulated; this number is equal to the size of the estimator memory for fixed memory methods, and equal to p+1 for growing memory methods.)

A. Fixed Memory

We define

$$\phi(i,j,n) = \sum_{k=n-M+p+1}^{n} s(k-i)\, s(k-j), \quad 0 \le i,j \le p, \tag{6}$$

where M>p is a finite constant. It is clear from (6) that $\phi(i,j,n)=\phi(j,i,n)$, i.e., $\Phi(n)$ is symmetric. Since the definition of $\Phi(n)$ given by (5) and (6) makes use of the signal samples s(n),s(n-1),...,s(n-M+1), the extent of the estimator memory is M samples. (One could also view the most recent M-p samples as representing the estimator memory, with the other p samples serving as initial conditions.) By a simple change of summation variable in (6), with r=k-1, it is easy to show that

$$\phi(i,j,n) = \phi(i-1,j-1,n-1), \quad 1 \le i,j \le p. \tag{7}$$

That is, the lower pxp submatrix of $\Phi(n)$ is identical to the upper pxp submatrix of $\Phi(n-1)$. Therefore, only the first row of $\Phi(n)$ has to be actually computed. (By symmetry, the first column is identical to the first row.) It can be easily shown from (6) that the elements of the first row, $\phi(0,j,n)$ are given by the recursive form:

$$\phi(0,j,n) = \phi(0,j,n-1) + s(n)\, s(n-j) - s(n-M+p)\, s(n-M+p-j), \quad 0 \le j \le p. \tag{8}$$

The number of multiplications required for computing $\Phi(n)$ by (7) and (8) is only 2(p+1). (Compare this with M(p+1) multiplications required in the computation of $\Phi(n)$ directly from signal samples using (6).) The covariance lattice method requires on the order of $(p^3+3p^2-4p)/2$ multiplications to solve for the parameters [1]. Therefore, the total number of multiplications per sample is about $(p^3+3p^2)/2$, most of this computational load being due to the covariance lattice solution. In terms of storage, the described BSE method requires an M-sample first-in first-out (FIFO) buffer for storing the most recent signal samples, and also a storage of size $(p+1)(p+2)/2$ for the elements of the symmetric covariance matrix.

The above approach can be easily extended to provide a new linear prediction estimate once every L signal samples instead of every sample. In other words, the M-sample analysis interval is shifted forward in time by L samples to obtain a new estimate.

B. Growing Memory

For the growing memory condition, we define the covariance as

$$\phi(i,j,n) = \sum_{k=p}^{n} s(k-i)\, s(k-j), \quad 0 \le i,j \le p, \tag{9}$$

where we have assumed that the signal sequence starts with s(0). The growing memory aspect of the estimator is obvious from (9), where the memory length is equal to n+1.

It is readily seen from (9) that

$$\phi(i,j,n) = \phi(i,j,n-1) + s(n-i)\, s(n-j), \quad 0 \le i,j \le p \tag{10}$$

Initially,

$$\phi(i,j,p) = s(p-i)\, s(p-j), \quad 0 \le i,j \le p. \tag{11}$$

The recursive computation of the covariance matrix elements in (10) requires $(p+1)(p+2)/2$ multiplications per signal sample. This, together with the covariance lattice solution, bring the total number of multiplications to about $(p^3+4p^2-p)/2$ for computing a new estimate. The amount of storage required is $p+1$ for storing the most recent samples, and $(p+1)(p+2)/2$ for storing the covariance matrix elements.

C. Fading Memory

By fading memory, we mean that recent data is given more emphasis than past data. This feature of "discounting" past data can be incorporated into either finite memory estimators or growing memory estimators. Since the introduction of fading in growing memory methods is straightforward, we consider that case first.

Growing Memory with Fading: Covariance computation in (10) may be modified to permit an exponential weighting of past data as follows:

$$\phi(i,j,n) = \beta\, \phi(i,j,n-1) + s(n-i)\, s(n-j), \quad 0 \le i,j \le p, \quad 0 \le \beta \le 1. \tag{12}$$

Notice that if $\beta=1$ (no fading), (12) becomes identical to (10); if $\beta=0$ (complete fading), we have fixed memory estimation ($M=p+1$ in (6)). With (11) still giving the initial covariance value, (12) can be rewritten as

$$\phi(i,j,n) = \sum_{k=p}^{n} \beta^{n-k}\, s(k-i)\, s(k-j), \quad 0 \le i,j \le p, \tag{13}$$

where the exponential weighting is explicitly shown.

Fixed Memory with Fading: For this case, inspection of (13) and (6) suggests a covariance definition that is the same as (13) except with the lower limit for the summation index k being $(n-M+p+1)$ instead of p. With this definition, (7) still holds; (8) is modified as follows:

$$\phi(0,j,n) = \beta\, \phi(0,j,n-1) + s(n)\, s(n-j) - \beta^{M-p}\, s(n-M+p)\, s(n-M+p-j), \quad 0 \le j \le p. \tag{14}$$

5. RECURSIVE ESTIMATION

In this section, we first extend the regular lattice approach to provide recursive estimation under each of the three memory conditions. Next, we apply Widrow's steepest descent least mean squares method to the lattice model given in Fig.1; the resulting estimator has a growing memory aspect which is different from that in the regular lattice approach.

A. Regular Lattice Approach

We shall consider Burg's method as an illustrative example in our discussion below. An important property of Burg's method that is not shared by other lattice methods, is that the estimate of the reflection coefficients results directly from the minimization of an error criterion [8,1]. The error is defined as the sum of the mean square values of the forward and backward residuals.

Referring to the lattice model in Fig.1, the memory at the input of stage $m+1$ at time n is represented by the residual sequences $\{f_m(k),\ b_m(k-1),\ k=n,n-1,\ldots\}$. The estimate of $K_{m+1}(n)$ is determined by minimizing the following error $E_{m+1}(n)$, which is the sum of forward and backward residuals at the output of that stage:

$$E_{m+1}(n) = \sum_{k}^{n} [f_{m+1}^2(k) + b_{m+1}^2(k)], \tag{15}$$

where the lower limit for the summation index is left unspecified to allow the use of either fixed memory or growing memory. Substituting (4b) and (4c) into (15), and equating the partial derivative of $E_{m+1}(n)$ with respect to K_{m+1} to 0, we obtain

$$K_{m+1}(n) = \frac{-2 \sum_{k}^{n} f_m(k)\, b_m(k-1)}{\sum_{k}^{n} [f_m^2(k) + b_m^2(k-1)]}. \tag{16}$$

(Notice that with the use of any other lattice method instead of Burg's method, expression (16) has to be appropriately modified.) The result in (16) is used to compute all the p reflection coefficients by substituting $m=0,1,\ldots,p-1$, in that

order. After a reflection coefficient at stage m is determined, the forward and backward residuals at the output of that stage are computed, so that K_{m+1} can then be obtained using (16). We have chosen to call the sequential procedure defined by (4) and (16) as recursive estimation since, as will be shown later, the expression for $K_{m+1}(n)$ in (16) can be rewritten as the sum of $K_{m+1}(n-1)$ and a correction term.

Defining

$$F_m(n) = \sum_k^n f_m^2(k), \tag{17a}$$

$$B_m(n) = \sum_k^n b_m^2(k-1), \tag{17b}$$

$$C_m(n) = \sum_k^n f_m(k)\, b_m(k-1), \tag{17c}$$

(16)becomes

$$K_{m+1}(n) = -2\, C_m(n)/[F_m(n) + B_m(n)]. \tag{18}$$

(Notice that the sum $F_m(n) + B_m(n)$ could have been defined as one quantity which is equal to the sum of the terms on the right hand sides of (17a) and (17b). This approach would reduce the storage required by the estimator, and as such should be preferred for actual implementation. However, we shall carry on the two residual norm squares in this section, as it allows one to think in terms of the "physical" signals at various nodes of the lattice shown in Fig.1.) The correlations in (17) can be computed recursively in time. Below we deal with this and other issues by considering each of the three memory conditions separately.

Fixed Memory: For this case, the lower limit for the summation index k in (15)-(17) is n-M+p+1, where M is the size of the estimator memory. With this lower limit in (17), we obtain

$$F_m(n) = F_m(n-1)+f_m^2(n)-f_m^2(n-M+p), \tag{19a}$$

$$B_m(n) = B_m(n-1)+b_m^2(n-1)-b_m^2(n-M+p-1), \tag{19b}$$

$$C_m(n) = C_m(n-1)+f_m(n)b_m(n-1) - f_m(n-M+p)\, b_m(n-M+p-1). \tag{19c}$$

Equations (4), (19) and (18) describe the sequential estimation method under consideration. Excluding the case M=p+1 (see below), the total number of computations per signal sample required by this method is 8p multiplications and p divisions if the M most recent samples of the residuals f_m and b_m, $0 \le m \le p-1$, are stored, or 5p multiplications and p divisions if the quantities f_m^2, b_m^2 and $f_m b_m$ (total = 3p x M) are stored instead. In both cases, the 3p correlations F_m, B_m and C_m have to be stored.

A special case of interest is obtained with M=p+1. For this case, each of the summations in (15)-(17) degenerates into a single term corresponding to k=n. In particular,

$$K_{m+1}^s(n) = \frac{-2\, f_m(n)\, b_m(n-1)}{[f_m^2(n) + b_m^2(n-1)]}. \tag{20}$$

where the superscript s denotes 'single term'. For this case, the estimate of the (m+1)th reflection coefficient at time n depends on the input residuals to that stage at that time only. Therefore, the updating relations (19) reduce to the middle term only in each equation, and hence are not explicitly required. The sequential procedure, described by (4) and (20), was suggested by Boll [4]; it requires 5p multiplications and p divisions per signal sample.

For the fixed memory condition, and excluding the degenerate case M=p+1, (16) can be rewritten in a recursive form as follows:

$$K_{m+1}(n) = K_{m+1}(n-1) + G_{m+1}(n).\, [y_{m+1}(n) - K_{m+1}(n-1)], \tag{21a}$$

$$G_{m+1}(n) = [v_{m+1}(n) - v_{m+1}(n-M+p)] \div V_{m+1}(n), \tag{21b}$$

$$v_{m+1}(k) = f_m^2(k) + b_m^2(k-1), \tag{21c}$$

$$V_{m+1}(n) = V_{m+1}(n-1) + v_{m+1}(n) - v_{m+1}(n-M+p), \tag{21d}$$

$$y_{m+1}(n) = [2\, f_m(n-M+p)\, b_m(n-M+p-1) - 2\, f_m(n)\, b_m(n-1)]/[v_{m+1}(n) - v_{m+1}(n-M+p)]. \tag{21e}$$

The quantity $y_{m+1}(n)$ that appears in the correction term in (21a) may be termed a "measurement" at time n. The gain term given in (21b) has an inverse relation to $V_{m+1}(=F_m+B_m)$, which is the sum of the norm

squares of the two input signals $f_m(k)$ and $b_m(k-1)$ defined over the memory span. Although the correction term in (21a) as described by (21b)-(21e) seems complicated, it is a function of only the quantities V_{m+1} and K_{m+1} at time n-1, and the input signals to stage m+1 at time instants n and n-M+p. While the recursive form (21) is useful for studying some of the properties of the estimation process, it should be cautioned that implementing the sequential procedure in that form is computationally less efficient than using (19) and (18).

Since most of the discussions presented for the fixed memory case, with simple modifications, apply to the growing memory and fading memory cases, we treat those cases below very briefly.

Growing Memory: With the lower limit for the summation index k in (15)-(17) equal to p for this case, we obtain

$$F_m(n) = F_m(n-1) + f_m^2(n), \tag{22a}$$

$$B_m(n) = B_m(n-1) + b_m^2(n-1), \tag{22b}$$

$$C_m(n) = C_m(n-1) + f_m(n)\, b_m(n-1). \tag{22c}$$

The growing memory recursive estimation method is thus described by (4), (22) and (18); it requires 5p multiplications and p divisions per signal sample, and needs to store 3p correlations given in (22).

For the growing memory condition, (16) can be rewritten in a recursive form that resembles the Kalman filter equation as follows:

$$K_{m+1}(n) = K_{m+1}(n-1) + G_{m+1}(n)\cdot [K^s_{m+1}(n) - K_{m+1}(n-1)], \tag{23a}$$

$$G_{m+1}(n) = v_{m+1}(n)/V_{m+1}(n), \tag{23b}$$

$$V_{m+1}(n) = V_{m+1}(n-1) + v_{m+1}(n), \tag{23c}$$

where v_{m+1} is defined by (21c). It is interesting to note that the estimate $K^s_{m+1}(n)$ produced by the degenerate case of the fixed memory estimator (see (20)) appears in the correction term in (23a) as a "measurement" at time n. Other comments that immediately follow (21) apply to (23) as well, with the major difference that the recursive form (23) is much simpler than (21).

The growing memory sequential estimation procedure has been used by Kang [3] and by Srinath and Viswanathan [12]. Both references, however, do not employ the error criterion (15). Following Itakura, Kang [3] uses (16) as an approximation to Itakura's PARCOR (partial correlation) coefficients, while reference [12] makes a stationarity assumption in deriving (16).

Fading Memory: In a manner analogous to that resulting in (13), a reasonable way to introduce fading is to weight the terms in the summation in (15) by β^{n-k}, $0 \le \beta \le 1$. It is easy to see that this weighting is carried over to (16) and (17). For recursively updating the correlations, (19) (or (22)) can be easily modified in the same way as we did for obtaining (12) (or (14)).

B. Steepest Descent Least Mean Squares (LMS) Approach

Widrow's "noisy" gradient LMS approach [13,14] has been used mainly for sequential estimation of the predictor coefficients. But, that method does not guarantee the stability of the all-pole model. Recently, Horvath applied Widrow's method to the lattice model of a pole-zero equalizer filter [15]. In the absence of zeroes, the model is as shown in Fig.1. Horvath used the lattice model primarily because checking the stability of the filter becomes a trivial problem (see (3)). Briefly, the recursive relations used in that LMS method are as follows:

$$K_m(n) = K_m(n-1) - \alpha_m(n)\, e(n)\cdot \left.\frac{\partial e(n)}{\partial K_m}\right|_{K_m = K_m(n-1)}, \quad 1 \le m \le p. \tag{24}$$

where $e(n)$ is the output residual $f_p(n)$ in Fig.1, and $\alpha_m(n)$ is a step-size parameter that is usually set to a small constant value, but that in general may be a function of time. The recursive form (24) is similar to (21) or (23) with the difference that the correction term is much simpler in (24); it is proportional to the negative of the instantaneous gradient of $e^2(n)$ with respect to K_m. (The procedure to compute this gradient is given in [15].)

With a fixed (i.e., data independent) step-size parameter sequence $\{\alpha_m(n)\}$, (24) can lead to filter instability. To overcome this problem, the step sizes may be changed whenever necessary to ensure that the updated reflection coefficients satisfy (3). This will guarantee the

filter stability at the expense of altering the nature of convergence of the LMS method.

From an inspection of (24) it follows that the LMS method has a growing memory. In view of the different correction terms in (22) and (24), it is evident that the growing memory aspect of the LMS method is different from that of the regular lattice method.

6. BLOCK SEQUENTIAL VERSUS RECURSIVE ESTIMATION

Some of the differences between the two estimation approaches were already stated in previous sections. Here, we emphasize two important differences.

First, for sample-by-sample estimation, block sequential methods are computationally more expensive than recursive methods. However, if estimates are not desired every sample, block methods can become more advantageous.

Second, the reflection coefficient estimates computed by the block sequential and recursive methods will be different, in general. To clarify this point, let us explain the operational details of the two approaches as follows (although specific implementations may not explicitly perform these operations). For each new signal sample, the block sequential approach uses all the signal samples in the memory to compute the estimate of the first reflection coefficient, passes them through the first stage in the lattice to generate the residuals $f_1(k)$ and $b_1(k)$ for all desired k, then computes the estimate of the second reflection coefficient from these residuals, etc. On the other hand, the recursive approach "ripples" each new signal sample, and only that sample, through the entire lattice to compute the estimate of all the reflection coefficients. Thus, the previous sample-by-sample (or instantaneous) estimates of the reflection coefficients determine the residuals which in turn are used for computing the current estimate. Therefore, only the estimates of the first reflection coefficient will be the same for both approaches; estimates of all other reflection coefficients will, in general, be different for the two approaches. Notice that the above operational description also indicates why block sequential methods are computationally more expensive than recursive methods.

Experimental comparisons of the results from the two estimation approaches will be presented at the conference.

7. APPLICATIONS IN SPEECH PROCESSING

Sequential estimation has been used in a number of speech processing applications. Some authors have dealt with sequential estimation of the predictor coefficients a_k in (2) [16-18]. Their methods do not guarantee the filter stability, unlike the methods presented in this paper. Below we briefly review the applications of sequential estimation methods in speech processing.

Determination of the instants at which certain speech events occur, such as glottal closure, may be accomplished through fixed memory sequential estimation [2]. If the ratio of mean-squared prediction error e(n) to mean-squared signal s(n) is used as a measure, and if the estimator memory is short compared with the pitch period, then the measure will often show sharp peaks whenever the time segment representing the estimator memory contains a glottal closure.

Sequential estimation has been used in pitch extraction schemes by Maksym [16] and Boll [4]. Boll reported that the estimation procedure given by (4) and (20) produced a spectrally flatter error sequence e(n) than block linear prediction methods did, with the fundamental frequency more clearly evident.

Next, we consider applications to efficient speech transmission systems. In fixed frame rate systems using block linear prediction, one set of p reflection coefficients (p≈12 for 10 kHz signal sampling rate) is computed for every data frame (typically 20 msec long), and transmitted to the receiver. Employing a sequential estimator that is initialized at the start of a data frame and terminated at the end of the data frame, one has as many sets of reflection coefficient estimates as there are speech samples in the data frame. Transmission of all of those estimates would tremendously increase the bit rate. One may select a best estimate in some sense and transmit that estimate only. Kang [3] has suggested transmitting the estimate that produces the minimum mean square value for the residual e(n). Use of this selection procedure with the growing memory sequential estimation method given by (4), (22) and (18) was found to reduce the "wobble" quality usually present in steady state regions of voiced sounds that are synthesized using block linear prediction methods [3]. Alternately, one may select the median of the estimates or mode of the probability histogram formed from the sample-by-sample estimates.

Other applications of sequential estimation include sequentially adaptive DPCM of speech [17], and potential use of sample-by-sample estimates in deciding transmission instances in variable frame rate systems [5,6].

ACKNOWLEDGMENT

This work was sponsored by the Information Processing Techniques branch of the Advanced Research Projects Agency.

REFERENCES

1. J. Makhoul, "New Lattice Methods for Linear Prediction," Proc. 1976 IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, pp. 462-465, April 1976.

2. H.W. Strube, "Determination of the Instant of Glottal Closure from the Speech Wave," J. Acoust. Soc. Amer., Vol.56, pp. 1625-1629, Nov. 1974.

3. G.S. Kang, Personal communications. (See, G.S. Kang, "Linear Predictive Narrowband Voice Digitizer," Proc.1974 EASCON Conf., Washington, D.C., pp.51-58, Oct. 1974, for a description of the hardware system.)

4. S. Boll, "Selected Methods for Improving Synthesis Speech Quality Using Linear Predictive Coding: System Description, Coefficient Smoothing and STREAK," UTEC-CSc-74-151, Comp. Science Dept., Univ. Utah, 1974.

5. R. Viswanathan and J. Makhoul, "Current Issues in Linear Predictive Speech Compression," Proc. 1974 EASCON Conf., Washington, D.C., pp. 577-585,Oct. 1974.

6. J. Makhoul, R. Viswanathan, L. Cosell and W. Russell, Natural Communication with Computers, Final Report, Vol. II, Speech Compression Research at BBN, Report No. 2976, Dec. 1974.

7. F. Itakura and S. Saito, "Digital Filtering Techniques for Speech Analysis and Synthesis," Proc. 7th Int. Cong. Acoust. (Budapest), 25-C-1, pp. 261-264, 1971.

8. J. Burg, "A New Analysis Technique for Time Series Data," NATO Advanced Study Institute on Signal Processing, Enschede, Netherlands, 1968. (See also M. Andersen, "On the Calculation of Filter Coefficients for Maximum Entropy Analysis," Geophysics, Vol.39, pp. 69-72, 1974.

9. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, Vol. 63, pp. 561-580, April 1975.

10. J.D. Markel and A.H. Gray, Jr., "Roundoff Noise Characteristics of a Class of Orthogonal Polynomial Structures," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-23, pp. 473-486, Oct. 1975.

11. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-23, pp. 309-321, June 1975.

12. M.D. Srinath and M.M. Viswanathan, "Sequential Algorithm for Identification of Parameters of an Autoregressive Process," IEEE Trans. Automatic Control, Vol. AC-20, pp. 542-546, Aug. 1975.

13. B. Widrow, "Adaptive Filters," in Aspects of Network and System Theory, R. Kalman and N. DeClaris, Eds. New York: Holt, Reinhart, and Winston, 1971, pp. 563-587.

14. L.J. Griffiths, "Rapid Measurement of Digital Instantaneous Frequency," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-23, pp. 207-222, April 1975.

15. S. Horvath, Jr., "Adaptive IIR Digital Filters for On-line Time-Domain Equalization and Linear Prediction," Presented at the IEEE Arden House Workshop on Digital Signal Processing, Harriman, N.Y., Feb. 1976.

16. J.N. Maksym, " Real-Time Pitch Extraction by Adaptive Prediction of the Speech Waveform," IEEE Trans. Audio and Electroacoust., Vol. AU-21, pp.149-154, June 1973.

17. J.D. Gibson, S.K. Jones and J.L. Melsa, "Sequentially Adaptive Prediction and Coding of Speech Signals," IEEE Trans. Communications, Vol. COM-22, pp. 1789-1797, Nov. 1974.

18. C. Schmid, "A Direct Method for Sequentially Updating Linear Predictor Coefficients for the Covariance Method," Proc. 1976 IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, pp. 479-480, April 1976.

APPENDIX 3

ADAPTIVE LATTICE METHODS FOR LINEAR PREDICTION

(Paper to be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK, April 1978.)

ADAPTIVE LATTICE METHODS FOR LINEAR PREDICTION

J. Makhoul and R. Viswanathan

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

ABSTRACT

A general method for adaptive updating of lattice coefficients in the linear predictive analysis of nonstationary signals is presented. The method is given as one of two sequential estimation methods, the other being a block sequential estimation method. The fast convergence of adaptive lattice algorithms is seen to be due to the orthogonalization and decoupling properties of the lattice. These properties are useful in adaptive Wiener filtering. As an application, a new fast start-up equalizer structure is presented. In addition, a one-multiplier form of the lattice is presented, which results in a reduction of computations.

1. INTRODUCTION

The lattice method of linear prediction was first introduced by Itakura [1,2] for speech analysis. A similar algorithm was proposed independently by Burg [3] in geophysics. Recently, Makhoul [4] showed the existence of a class of lattice methods of which the methods of Itakura and Burg are special cases. All these methods guarantee the stability of the corresponding all-pole filter, with or without windowing of the signal, independently of the stationarity properties and duration of the signal, and with finite wordlength computations. Also, for data compression purposes, quantization of the reflection coefficients may be accomplished within the lattice recursion. In addition, Makhoul [4] developed the so-called covariance-lattice methods, which compute the lattice model parameters from the covariance of the signal, with a 3-4 fold saving in computation over the methods of Itakura and Burg.

The only known disadvantage of lattice methods is that, if the signal is not windowed, the computed model parameters may not minimize the output mean-square error, resulting in a suboptimal solution [4]. For most applications, this disadvantage is of no consequence.

In addition to the advantages given above, the lattice has a most important orthogonalization property: the "decoupling" of consecutive stages of the lattice. This property substitutes the global minimization at the lattice output with a sequence of local minimization problems, one at each stage of the lattice.

This paper explores the adaptive estimation of lattice parameters in a nonstationary environment. Earlier related work may be found in [5-7]. Griffiths [7] pointed out that the decoupling property results in a convergence that is independent of the signal. In this paper we show how the orthogonalization property may be used to great advantage in adaptive Wiener filtering. In particular, we present as an application a new fast start-up adaptive equalizer. Adaptive lattice methods promise to be useful in areas where transversal, predictive, or finite impulse response (FIR) filters are used in an adaptive manner.

2. LATTICE PRELIMINARIES

Fig. 1 shows the basic two-multiplier lattice of Itakura and Saito [1,2]. From Fig. 1, the following relations hold:

$$f_0(n) = g_0(n) = s(n) \tag{1a}$$

$$f_m(n) = f_{m-1}(n) + K_m g_{m-1}(n-1) \tag{1b}$$

$$g_m(n) = K_m f_{m-1}(n) + g_{m-1}(n-1). \tag{1c}$$

$x(n)$ is the input signal, $f_m(n)$ is the "forward" residual at stage m, $g_m(n)$ is the "backward" residual, and K_m is the reflection coefficient. Let the forward transfer function up to stage m be

$$A_m(z) = \sum_{k=0}^{m} a_m(k)\, z^{-k}, \quad a_m(0) = 1, \tag{2}$$

where $a_m(k)$ are the predictor coefficients for an mth order predictor. Then the backward transfer function up to stage m is given by (2) with the order of the coefficients reversed. The predictor coefficients are computed from the reflection coefficients using the recursion

$$a_m(m) = K_m$$

$$a_m(k) = a_{m-1}(k) + K_m a_{m-1}(m-k), \quad 1 \le k \le m-1. \tag{3}$$

The stability of the all-pole filter $1/A_p(z)$ is guaranteed iff

$$|K_m| < 1, \quad 1 \le m \le p. \tag{4}$$

CH1285-6/78/0000-0083$00.75@1978IEEE

3. SEQUENTIAL ESTIMATION METHODS

A sequential estimation method, by our definition, provides a new estimate for the reflection coefficients at each time instant n. We differentiate two types of sequential estimation methods [6]:

(1) Block estimation,
(2) Adaptive estimation.

Block estimation is the usual method of linear prediction analysis, where one value of each reflection coefficient is estimated for a whole block of data. The analysis is repeated over again as each signal sample is added to the block of data. In contrast to block estimation, adaptive estimation determines a new estimate at time n as a function of the last estimate at time n-1 and a "measurement" at time n. Below, we present both types of estimation methods.

4. BLOCK ANALYSIS OF NONSTATIONARY SIGNALS

We assume that x(n) is a nonstationary signal, and that we wish to estimate the reflection coefficients K_m at each instant of time n. We shall take advantage of the decoupling property of the lattice (even though it is only approximately true in the nonstationary case), and determine each $K_m(n)$ by minimizing some function of the forward and backward residual energies at that stage. Furthermore, since in a time-varying situation we are mainly interested in the most recent history of the signal, it is reasonable to weight the residuals such that the more recent values are given greater importance. We are thus led to minimizing a mean-square type of error of the form:

$$E_m(n) = \sum_{k=-\infty}^{n} w(n-k)\, e_m^2(k) \qquad (5)$$

where w(n) is the weighting sequence, or window, and $e_m^2(k)$ is the residual energy at time k. We shall have more to say about the window later on. As for the residual energy, we shall consider here only one case, the sum of the forward and backward residual energies:

$$e_m^2(k) = f_m^2(k) + g_m^2(k). \qquad (6)$$

Substituting (6) and (1) in (5) and minimizing $E_m(n)$ with respect to $K_m(n)$ results in:

$$K_m(n) = -\frac{2\sum_{k=-\infty}^{n} w(n-k)\, f_{m-1}(k)\, g_{m-1}(k-1)}{\sum_{k=-\infty}^{n} w(n-k)\,[f_{m-1}^2(k) + g_{m-1}^2(k-1)]} \qquad (7a)$$

$$= -\frac{C(n)}{D(n)} \qquad (7b)$$

The value of K as given by (7) is always guaranteed to obey (4). Other possibilities exist for defining K such that (4) is guaranteed [4], but they will not be discussed here.

In the block method, $K_1(n)$ is computed first, using (7) and the input signal (see(1a)). Then, the residuals $f_1(k)$ and $g_1(k)$ are computed using (1) for <u>all</u> time up to n. Then, $K_2(n)$ is computed from (7), followed by the computation of $f_2(k)$ and $g_2(k)$, and so on for all stages. The whole process is then repeated at time n+1, with the residuals having to be completely reevaluated for all time up to n+1. The amount of computation is clearly large, and so is the apparent amount of storage. However, one can effect substantial savings in both by using the covariance lattice method instead [4], and by recursively updating the signal covariance [6]. It is important to note that, in the block method, the value of $K_m(n+1)$ does <u>not</u> depend in any simple way on $K_m(n)$. This is to be contrasted with the adaptive method described in Section 5.

Windowing

We point out at the outset that windowing of the error in (5) is very different from windowing of the signal. Windowing the signal results in a stationary signal, while windowing the error does not affect the stationarity of the signal; it merely weights the different error values. Signal or data windows may be quite arbitrary, and may take on positive and negative values. In contrast, the error window in (5) must be always nonnegative. In particular, we must have

$$w(n) \geq 0, \quad n \geq 0,$$
$$w(n) = 0, \quad n < 0. \qquad (8)$$

Negative values are not allowed since they will result in cancellation of errors, which is generally undesirable.

As examples, we shall give one FIR window and one recursive window. The FIR window is the usual rectangular window of width M:

$$w_1(n) = 1, \quad 0 \leq n \leq M-1,$$
$$= 0, \text{ otherwise.} \qquad (9)$$

This window has some bad effects as a signal window but has good properties as an error window. The recursive window is the impulse response of a single real pole:

$$w_2(n) = \beta^n, \quad n \geq 0, \quad 0 < \beta < 1$$
$$= 0, \quad n < 0. \qquad (10)$$

From (10) and (7), one can compute C(n) and D(n) recursively using

$$C(k) = \beta C(k-1) + 2 f_{m-1}(k) g_{m-1}(k-1) \qquad (11a)$$

$$D(k) = \beta D(k-1) + f_{m-1}^2(k) + g_{m-1}^2(k-1) \qquad (11b)$$

for all k up to n. Other recursive windows may be defined, but because of condition (8), all such windows must be the impulse responses of lowpass filters with <u>positive real</u> poles.

5. ADAPTIVE ESTIMATION

In adaptive estimation, we assume given $K_m(n)$, $1 \le m \le p$, at time n, and the forward and backward residuals up to time n. The problem is then to estimate $K_m(n+1)$, $1 \le m \le p$, at time n+1 using the given quantities. We shall employ the estimate in (7) but in a different manner:

$$K_m(n+1) = -\frac{C(n)}{D(n)} \quad . \tag{12}$$

Given $K_m(n)$ and $g_m(n-1)$, $1 \le m \le p$, one computes $f_m(n)$ and $g_m(n)$ for all stages using (1). Then $K_m(n+1)$, $1 \le m \le p$, are computed from (12), and so on. In contrast with the block method, the residuals are computed only once for each point in time.

The windows $w_1(n)$ and $w_2(n)$ may also be used in adaptive estimation. For example, with $w_2(n)$ one can use (11) with k=n. It is clear from the recursive computation in (11) that only 6 multiplications (neglecting multiplication by 2) and 1 division are needed to compute each of the reflection coefficients at each point in time. In addition, the necessary memory is minimal. The rectangular window in (9) requires 2 fewer multiplications, but in exchange requires memory proportional to M, the window width. Therefore, the main advantage of adaptive estimation over the block method is the reduced computation, and the reduced storage when using the recursive window. The price to be paid is that adaptive estimation is noisier; we view the adaptive estimation method as an approximation to the block method. Examples illustrating the difference between the two methods will be given in the conference.

An algorithm using $w_2(n)$ was used by Itakura in his original hardware realization of the lattice in a speech vocoder system [8]. A similar vocoder has been designed by Kang [9].

LMS Interpretation

Using $w_2(n)$ and therefore (11), one can show that $K_m(n+1)$ for this special window may be written as an update on $K_m(n)$:

$$K_m(n+1) = K_m(n) - \frac{f_{m-1}(n)g_m(n) + g_{m-1}(n-1)f_m(n)}{D(n)} \quad , \tag{13}$$

where D(n) is given recursively by (11b) with k=n. For the special case $\beta = 1$, D(n) increases continuously and the correction term in (13) tends to zero as n goes to infinity. In this case K_m tends to its optimal value with probability 1, assuming a stationary signal. For $\beta < 1$, one can show that (13) becomes identical to the LMS lattice estimate of Griffiths [7] with a step size $\alpha = 1-\beta$.

Kalman Filter Interpretation

One can show [6] that (13) can be rewritten in the form of a Kalman filter:

$$K_m(n+1) = K_m(n) + G_m(n+1)[K_m^a(n+1) - K_m(n)] \tag{14}$$

where

$$K_m^a(n+1) = -\frac{2f_{m-1}(n)g_{m-1}(n-1)}{f_{m-1}^2(n) + g_{m-1}^2(n-1)} \tag{15}$$

can be viewed as a single "measurement" at time n+1, and $G_m(n+1)$ is a gain term at n+1 given by:

$$G_m(n+1) = \frac{d_m(n)}{D_m(n)} \tag{16}$$

where

$$d_m(n) = f_{m-1}^2(n) + g_{m-1}^2(n-1) \tag{17a}$$

and

$$D_m(n) = \beta D_m(n-1) + d_m(n). \tag{17b}$$

$d_m(n)$ may be interpreted as the instantaneous residual variance, while $D_m(n)$ is the total variance.

6. ONE-MULTIPLIER LATTICE

The two-multiplier lattice in Fig. 1 is only one of many possible lattice implementations of the all-zero forward and backward prediction filters. Some of the implementations have a single multiplier, which would be useful if a smaller number of multiplies is desired. Fig. 2 shows one such implementation. Others may be found in [10].

7. A NEW FAST START-UP EQUALIZER

As an application to the adaptive lattice we propose a new fast start-up equalizer. This will be useful in polling applications where the initial time for the adaption process is desired to be as small as possible. Chang [11] proposed an equalizer structure that reduces the start-up time drastically. The general form of the structure is shown in Fig. 3. The tap coefficients c_i are adjusted such that the mean-square error between y(n) and some reference signal is minimized. If the filters are selected such that the signals $z_i(n)$ are orthonormal, then the tap coefficients c_i can be adjusted to their optimum values in one step [11].

In the specific equalizer proposed by Chang, the filter signals $z_i(n)$ are obtained from x(n) by a linear transformation $\underline{z} = \underline{P}\,\underline{x}$, where $\underline{z} = [z_1(n) \ldots z_N(n)]^T$, $\underline{x} = [x(n) \ldots x(n-N+1)]^T$, and $\underline{P}$ is an NxN transformation matrix that obeys

$$\underline{P}\,\underline{R}\,\underline{P}^T = \underline{I}, \tag{18}$$

where $\underline{R}$ is the NxN autocorrelation matrix of the signal x(n). The signal x(n) is taken here to be the impulse response of the channel. The solution chosen for $\underline{P}$ by Chang is $\underline{P} = \underline{D}^{-1/2}\,\underline{Q}^T$, where $\underline{R} = \underline{Q}\,\underline{D}\,\underline{Q}^T$, $\underline{Q}$ is a matrix whose columns are the orthonormal eigenvectors of $\underline{R}$, and $\underline{D}$ is a diagonal matrix whose elements are the eigenvalues of $\underline{R}$. However, the use of $\underline{P}$ requires N^2 coefficients with an equal number of multiplies, which can become excessive for large N. Below, we give our lattice structure for the fast start-up equalizer, where the number of coefficients is only a linear function of N.

Instead of an eigenvector decomposition for $\underline{R}^{-1}$ one can perform an $\underline{L}\,\underline{D}\,\underline{L}^T$ decomposition, where $\underline{L}$ is a lower triangular matrix, using the Gram-Schmidt

orthogonalization process. The transformation $\underline{L}\,\underline{x}$ turns out to be the sequence of backward residuals $g_i(n)$, which are orthogonal to each other. In particular, we have [10]

$$\overline{g_i(n)\, g_j(n)} = \begin{cases} E_i\,, & i=j\,, \\ 0\,, & i\neq j\,, \end{cases} \tag{19}$$

where E_i is the minimum residual energy at the ith stage. In order to render the orthogonal signals $g_i(n)$ orthonormal, one needs to multiply $g_i(n)$ by $E_i^{-1/2}$. Hence, $\underline{P} = \underline{E}^{-1/2}\underline{L}$ results in the desired solution. One may normalize with respect to the signal energy and have a transformation $z_i(n) = V_{i-1}^{-1/2}\, g_{i-1}(n)$, where

$$V_i = \prod_{m=1}^{i} (1 - K_m^2)$$

for the two-multiplier lattice of Fig. 1, or

$$V_i = \prod_{m=1}^{i} \frac{1+K_m}{1-K_m}$$

for the one-multiplier lattice of Fig. 2. The final equalizer structure is shown in Fig. 4. The total number of extra coefficients employed is 3(N-1) for the two-multiplier lattice or 2(N-1) for the one-multiplier lattice. The lattice is adapted to the channel first, as described in Section 5, and the tap coefficients should then adapt in one step.

ACKNOWLEDGMENTS

This work was sponsored by the Information Processing Techniques Office of the Advanced Research Projects Agency under Contract No. MDA903-75-C-0180.

REFERENCES

1. F. Itakura and S. Saito, "Digital Filtering Techniques for Speech Analysis and Synthesis," Proc. 7th Int. Cong. Acoust., Budapest, Paper 25-C-1, pp. 261-264, 1971.
2. F. Itakura and S. Saito, "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," IEEE Conf. on speech comm. and Processing, Newton, Mass., pp. 434-437, April 1972.
3. J.P. Burg, "Maximum Entropy spectral Analysis," Ph.D. dissertation, Geophysics Dept., Stanford Univ., Stanford, CA, May 1975.
4. J. Makhoul, "Stable and Efficient Lattice Methods for Linear Prediction," IEEE Trans. Acoust., Speech, Signal Processing, pp. 423-428, Oct. 1977.
5. M.D. Srinath and M.M. Viswanathan, "Sequential Algorithm for Identification of Parameters of an Autoregressive Process," IEEE Trans. Automatic control, pp. 542-546, Aug. 1975.
6. R. Viswanathan and J. Makhoul, "Sequential Lattice Methods for Stable Linear Prediction," EASCON '76, Washington, D.C., paper 155, 1976.
7. L.J. Griffiths, "A Continuously-Adaptive Filter Implemented as a Lattice Structure," IEEE Int. Conf. Acoust., Speech, Signal Processing, Hartford, Conn., pp. 683-686, May 1977.
8. F. Itakura, personal communication.
9. G.S. Kang, personal communication. (See, G.S. Kang, "Linear Predictive Narrowband Voice Digitizer," Proc. 1974 EASCON Conf., Washington, D.C., pp.51-58, Oct. 1974, for a description of the hardware system.)
10. J. Makhoul, "A Class of All-Zero Lattice Digital Filters: Properties and Applications," submitted to IEEE Trans. Acoustics, Speech and Signal Processing, Sept. 1977.
11. R.W. Chang, "A New Equalizer Structure for Fast Start-up Digital Communication," Bell Syst. Tech. J., pp. 1969-2014, July-Aug. 1971.

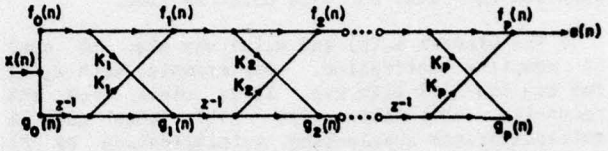


Fig. 1 Basic all-zero lattice filter.

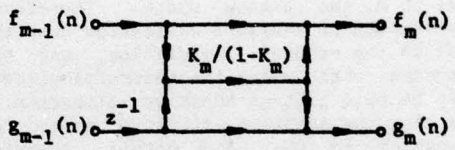


Fig. 2 The mth stage of a one-multiplier lattice.

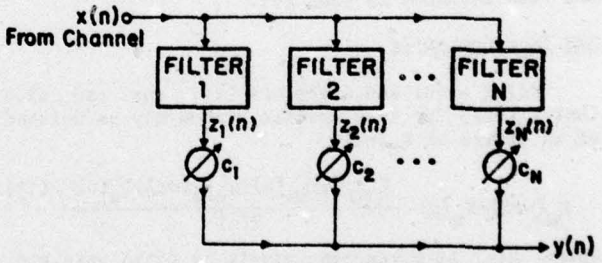


Fig. 3 Generalized equalizer structure.

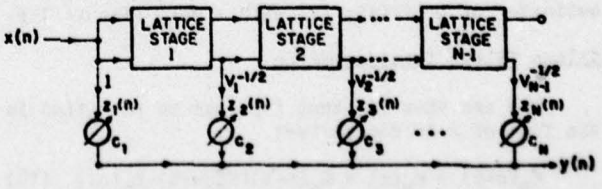


Fig. 4 Lattice fast start-up equalizer structure.

APPENDIX 4

METHODS FOR NONLINEAR SPECTRAL DISTORTION OF SPEECH SIGNALS

(Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, April 1976.)

METHODS FOR NONLINEAR SPECTRAL DISTORTION OF SPEECH SIGNALS

John Makhoul

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

The spectral distortion of speech signals, without affecting the pitch or the speed of the signal, has met with some difficulty due to the need for pitch extraction. This paper presents a general analysis-synthesis scheme for the arbitrary spectral distortion of speech signals without the need for pitch extraction. Linear predictive warping, cepstral warping, and autocorrelation warping, are given as examples of the general scheme. Applications include the unscrambling of helium speech, spectral compression for the hard of hearing, bit rate reduction in speech compression systems, and efficiency of spectral representation for speech recognition systems.

1. Introduction

Arbitrary spectral distortion of any finite sampled signal can be easily accomplished by computing the discrete Fourier transform (DFT) of the signal, performing the desired spectral distortion, and then taking the inverse DFT. (The resulting signal is an approximation to the desired spectrally distorted signal in the same measure as the DFT is an approximation to the z transform. Arbitrary accuracy can be achieved by increasing the order of the DFT.) In applying this method to the spectral distortion of voiced speech signals, the spectral envelope is distorted as well as the voicing (pitch) characteristics. For many applications, the distortion is usually desired for the spectral envelope, but not for the pitch. Thus it becomes necessary to separate the pitch (source) information, distort the spectral envelope, and then resynthesize using the extracted source information.

Certain existing research systems [1-3] for the nonlinear spectral distortion of speech signals separate the source information by making voiced/unvoiced decisions and performing pitch extraction. A different approach was taken by Suzuki et al. [4] for the unscrambling of helium speech, where pitch extraction was not used. In their work, the source information was obtained as the residual signal in a linear predictive analysis of the speech signal. The spectral distortion was performed in the time domain on the impulse response of the all-pole filter. However, the only type of distortion attempted was a linear one, and it was effected by interpolation in the time domain. In this paper we describe a general analysis-synthesis system for the nonlinear spectral distortion of speech signals, without the need for pitch extraction. The generality of the system is achieved by performing the spectral distortion directly in the frequency domain. Three methods, linear predictive warping, cepstral warping, and autocorrelation warping, are given as examples of the general scheme.

2. General System

The general analysis-synthesis system for spectral distortion is shown in Fig. 1. The speech signal s(n) is passed through a filter whose magnitude frequency response is the inverse of the envelope of the signal spectrum. The output of the inverse filter is the residual signal e(n), which contains mainly the source information. Since all the resonant structure of the signal is removed by the inverse filter, e(n) will have an essentially flat spectral envelope. The residual signal is then used as input to a synthesis filter whose magnitude frequency response is equal to the desired distorted or warped spectral characteristics. The output of the synthesis filter, s′(n), is then the transformed signal with the same source characteristics as s(n), but with a spectrum that is a distorted version of the spectrum of s(n).

One important property of the system in Fig. 1 is that the sampling rate remains fixed throughout the system. If for some reason the sampling rate at the output is desired to be different from that at the input, then one needs to perform down sampling on the residual signal, or else perform pitch extraction.

There remains the specification of the inverse filter and synthesis filter parameters. This is described next.

3. Nonparametric Warping

The dashed box in Fig. 1 shows the general scheme for spectral warping and for the specification of the inverse and synthesis filter parameters. A more detailed block diagram is shown in Fig. 2. The spectrum $P(\omega)$ of the signal s(n) is computed by windowing the signal and taking the magnitude squared of its Fourier transform. $P(\omega)$ is then smoothed to retain the requisite resonant structure. The smoothed spectrum $\hat{P}(\omega)$ is then inverted. The resulting inverse smoothed spectrum is then used to determine the impulse response a(n) of the inverse filter A(z). Assuming a minimum phase implementation, a(n) can be computed from $\hat{P}^{-1}(\omega)$ through the use of the cepstrum. Details can be found in Oppenheim and Schafer [5].

The impulse response h(n) of the synthesis filter H(z) can be computed using the lower branch in Fig. 2. The signal spectrum is distorted then smoothed to obtain $\hat{P}'(\omega)$. Again, assuming a minimum phase implementation, h(n) can be computed from $\hat{P}'(\omega)$ using the cepstral method. An alternative method to compute $\hat{P}'(\omega)$ is shown by the dashed lines in Fig. 2, where the smoothed spectrum $\hat{P}(\omega)$ is directly distorted. Note that the two alternative methods do not result in identical spectra for $\hat{P}'(\omega)$. Which method to use depends on the particular application.

Since the method to compute the minimum phase impulse response from a spectrum involves taking the DFT, it is desirable for efficiency purposes to have the frequency values in the spectrum be equally spaced and their number be an integral power of 2, so that one can make use of the FFT. If $P(\omega)$ is computed using the FFT, then the two conditions can be easily met for $\hat{P}^{-1}(\omega)$. However, because of the spectral distortion in the lower branch of Fig. 2, the spectral values of $\hat{P}'(\omega)$ will not be equally spaced in general. By simple interpolation in the frequency domain, the spectral values can be computed at equally spaced frequencies, thus opening the way to the use of the FFT.

The smoothing of the signal spectrum to obtain the spectral envelope can be done in many different ways. Here, we give two popular nonparametric methods which comprise two of the three methods of spectral warping that are presented in the paper. A parametric method is given in the next section.

Autocorrelation Warping

In this method the spectrum is smoothed by applying a window to the autocorrelation. This is the well-known method of spectral estimation used by statisticians [6].

Cepstral Warping

In this method the log spectrum is smoothed by applying a window to the cepstrum. This method of spectral smoothing has been used extensively in speech analysis [5].

4. Linear Predictive Warping

We have called the types of spectral warping in the previous section "nonparametric" because no specific model is used to determine the impulse response of the inverse and synthesis filters. In this section we use the all-pole linear prediction model as a basis to determine the parameters of the two filters. Fig. 3 shows a schematic diagram of linear predictive (LP) warping. The parameters $a(k)$ of the inverse filter are simply the predictor coefficients which are obtained by spectral LP [7] as a solution to the set of linear equations:

$$\sum_{k=1}^{p} a(k)\, R(i-k) = -R(i), \quad 1 \le i \le p, \qquad (1)$$

where p is the number of poles in the model, and $R(i)$ is the autocorrelation of the signal, which can be computed either by taking the FFT of the spectrum $P(\omega)$, or directly from the signal. Note that the method of spectral LP inherently smoothes the signal spectrum, with the degree of smoothing being controlled by the number of poles p. Referring to Fig. 2, the smoothed spectrum $\hat{P}(\omega)$ in this case is given by the all-pole model spectrum:

$$\hat{P}(\omega) = \frac{1}{\left|1+\sum_{k=1}^{p} a(k)e^{-jk\omega}\right|^2}. \qquad (2)$$

The parameters $a'(k)$ of the synthesis filter are obtained as a solution to a set of equations analogous to (1) with $a(k)$, $R(i)$ and p replaced by $a'(k)$, $R'(i)$ and q, respectively, where $R'(i)$ is the Fourier transform of the distorted spectrum $P'(\omega)$, and q is the number of poles in the synthesis filter. In general, $q \neq p$, and its choice depends on the application.

The parameters $a(k)$ need not be computed using spectral LP, which is essentially equivalent to the autocorrelation method of LP. Instead, one could use the covariance, lattice or covariance lattice methods [8]. In that case, $P(\omega)$ is undefined. So following the dashed line in Fig. 2, we compute $\hat{P}(\omega)$ from (2), distort it, then apply spectral LP to the resulting distorted spectrum in order to evaluate the coefficients $a'(k)$ of the synthesis filter.

5. Applications

There are many possible applications for the methods of nonlinear spectral warping given above. Below, we shall give four applications: two of these use the spectral warping for a more efficient representation of the spectrum, and two are analysis-synthesis systems for generating speech that is spectrally distorted.

Efficiency of Spectral Representation

In applications such as speech recognition and speech compression, it is more important to represent the spectrum accurately at low frequencies (<3 kHz) than at high frequencies (>3 kHz). Normally, anywhere between 17-20 poles are needed for an all-pole LP representation of speech spectra with a bandwidth of 7.5 kHz (sampling frequency 15 kHz). Using LP warping, for example, with frequencies above 3 kHz being heavily warped, one could have a good representation using only 12-14 poles. In this manner, one could still perform accurate formant extraction for the first three formants, with the higher formants being represented by wide spectral peaks, which is all that is usually needed [9].

For speech compression, this enables one to have wide-band, high quality speech at low bit rates, since fewer coefficients need to be transmitted. This idea has been recently implemented in an LPCW vocoder: an LPC vocoder with spectral warping [10].

Unscrambling of Helium Speech

In order to render speech spoken in a helium-oxygen mixture more intelligible, it is necessary to compress the bandwidth from about 12 kHz down to 5 kHz. In addition to this linear warping, one might need to perform additional nonlinear warping at low frequencies to compensate for high pressure effects [1,2,4]. Heretofore, such nonlinear warping had not been possible.

Since the bandwidth is reduced to 5 kHz, one must still define values for the spectrum between 5 and 12 kHz (assuming a 24 kHz sampling frequency). The reason is that in our analysis-synthesis system the sampling

rate remain fixed. It is usually sufficient to assign a positive constant for the spectrum between 5 and 12 kHz that is a fixed number of decibels below the maximum value in the spectrum. A value of zero, however, is not recommended.

Speech for the Hard of Hearing

Many people with severe hearing loss cannot hear frequencies much above 1 kHz [11]. An idea that some researchers have had is to compress the speech spectrum so that the most important part of the spectrum (up to 3 kHz) is compressed down to less than 1 kHz. It is hoped that this squeeze of the spectral information into a small bandwidth would aid the hard of hearing in listening to speech, and would eventually lead to the design of more effective hearing aids. It is easy to show that a simple linear compression of the spectrum to less than 1 kHz is quite unintelligible. However, the results improve dramatically if a nonlinear warping that emphasizes low frequencies is effected.

The technical details for this application are very similar to those described above for the unscrambling of helium speech.

6. Conclusion

A general analysis-synthesis system for the nonlinear spectral distortion of speech signals was described. The method does not need any pitch extraction, and allows for the arbitrary specification of the warping function. The latter is accomplished by performing the warping directly in the frequency domain. Depending on the type of spectral smoothing used, three methods resulted: autocorrelation, cepstral and linear predictive warping. Applications for these methods included bit rate reduction in high quality speech compression systems, efficient spectral representation for use in speech recognition systems, unscrambling of helium speech, and spectral compression for the hard of hearing.

Acknowledgment

The author thanks Lynn Cosell for implementing the LP warping algorithm for use in the hard of hearing application.

References

1. F. Quick, "Helium Speech Translation Using Homomorphic Techniques," S.M. Thesis, M.I.T., Cambridge, Mass., June 1970.

2. V. Zue, "Translation of Divers' Speech Using Digital Frequency Warping," QPR No. 101, R.L.E., M.I.T., Cambridge, Mass., 175-182, April 1971.

3. A. Oppenheim and D. Johnson, "Discrete Representation of Signals," Proc. IEEE, Vol. 60, 681-691, June 1972.

4. H. Suzuki, G. Ooyama and K. Kido, "Analysis-Conversion-Synthesis System for Improving Naturalness and Intelligibility of Speech at High-Pressure Helium Gas Mixture," Preprints, Speech Communication Seminar, Stockholm, Vol. 1, 97-105, Aug. 1974.

5. A. Oppenheim and R. Schafer, Digital Signal Processing, Ch. 10, New Jersey: Prentice-Hall, 1975.

6. R. Blackman and J. Tukey, The Measurement of Power Spectra, New York: Dover, 1958.

7. J. Makhoul, "Spectral Linear Prediction: Properties and Applications," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 283-296, June 1975.

8. J. Makhoul, "New Lattice Methods for Linear Prediction," IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, April 1976.

9. S. Itahashi and S. Yokoyama, "A Method of Formant Extraction Utilizing the Mel Scale," J. Acoust. Soc. Japan, Vol. 30, No. 12, 677-678, Dec. 1974.

10. J. Makhoul and L. Cosell, "LPCW: An LPC Vocoder with Linear Predictive Spectral Warping," IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, April 1976.

11. J.D. Schein and M.T. Delk, Jr., The Deaf Population of the United States, National Association of the Deaf, Silver Spring, Maryland, 1974.

(Figures on next page)

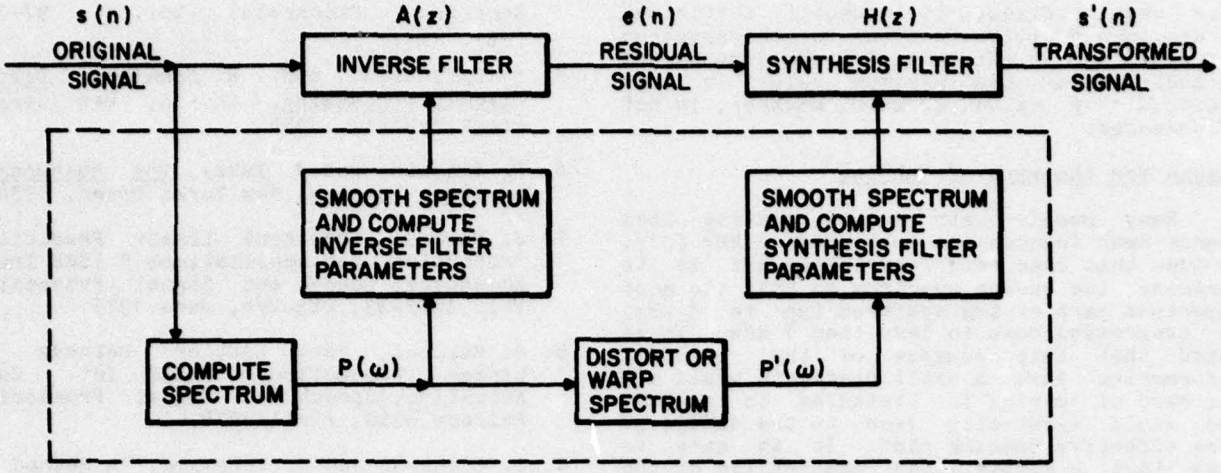


Fig. 1. General analysis-synthesis system for spectral warping.

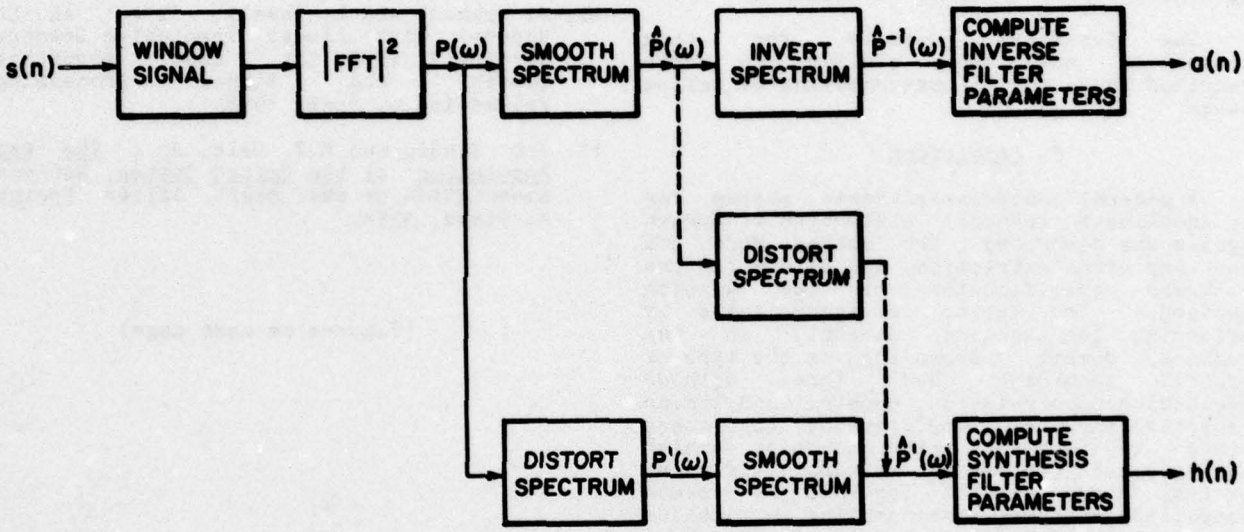


Fig. 2. Computation of the inverse and synthesis filter parameters.

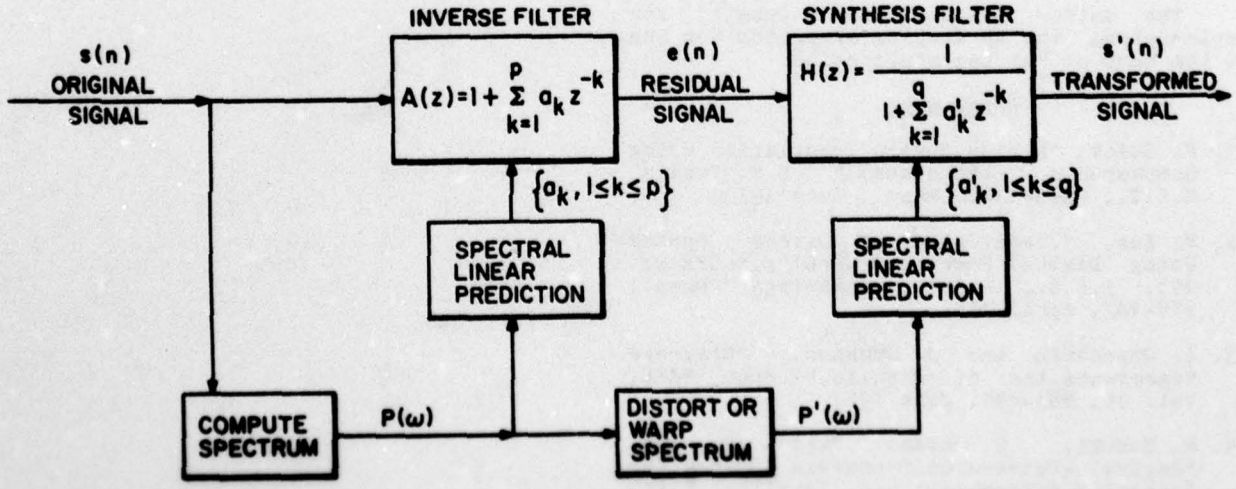


Fig. 3. Analysis-synthesis system for linear predictive warping.

APPENDIX 5

LPCW: AN LPC VOCODER WITH LINEAR PREDICTIVE SPECTRAL WARPING

(Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadephia, PA, April 1976.)

LPCW: AN LPC VOCODER WITH LINEAR PREDICTIVE SPECTRAL WARPING

John Makhoul
Lynn Cosell

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

In ordinary linear prediction the speech spectral envelope is modeled by an all-pole spectrum. The error criterion employed guarantees a uniform fit across the whole frequency range. However, we know from speech perception studies that low frequencies are more important than high frequencies for perception. Therefore, a minimally redundant model would strive to achieve a uniform perceptual fit across the spectrum, which means that it should be able to represent low frequencies more accurately than high frequencies. This is achieved in the LPCW vocoder: an LPC vocoder employing our recently developed method of linear predictive warping (LPW). The result is improved speech quality for the same bit rate.

1. Introduction

Narrow-band LPC vocoders with transmission rates less than 4800 bps have generally dealt with speech sampled at less than 10 kHz and usually closer to 6.5 kHz. Since the bit rate needed for transmission is roughly proportional to the sampling rate, it is argued justifiably that the possible increase in speech intelligibility and quality in going to 10 kHz is not commensurate with the increase in bit rate, and so sampling rates closer to 6.5 kHz have dominated the vocoder scene. The argument can also be phrased another way. If the bit rate is to remain fixed (e.g., 2400 bps), then an increasing the number of bits for each frame means that one is forced to transmit fewer frames per second. Thus, while spectral fidelity is increased for each transmitted frame, the accuracy in following the dynamic aspects of the signal is decreased.

Traditional channel vocoder systems have solved this problem by positioning their filters nonlinearly such that more filters are at low frequencies than at high frequencies [1]. It is not unusual to see a filter placed as high as 7 kHz in a channel vocoder. Thus, the total speech bandwidth represented can be about 7 kHz, which is to be contrasted with bandwidths closer to 3 kHz in LPC vocoders. (It should not be concluded from this, though, that channel vocoders produce higher quality speech than LPC vocoders for a given bit rate.)

A hybrid solution was introduced in the TRIVOC vocoder [2], which used an LPC representation at low frequencies and a channel vocoder at higher frequencies. This, of course, has the disadvantage of having to program two different vocoder systems.

This paper presents LPCW: an LPC vocoder that is capable of representing low frequencies better than high frequencies. This suggests the possibility of wide-band speech at low bit rates.

2. Linear Predictive Warping

The idea behind LPCW is quite simple: Warp the spectrum such that high frequencies are compressed relative to low frequencies, then apply spectral linear prediction [3] to the warped spectrum. Because the resulting representation is uniform across the warped spectrum, it means that low frequencies are better matched than higher frequencies since the latter are compressed.

The procedure for computing the coefficients of the warped spectrum is as follows:

(a) Window the signal and compute its spectrum.
(b) Warp the spectrum as desired.
(c) Take the Fourier transform of the warped spectrum to get the autocorrelation R(i).
(d) Solve for the predictor parameters from the normal equations:

$$\sum_{k=1}^{p} a(k)\, R(i-k) = -R(i), \quad 1 \le i \le p, \qquad (1)$$

where a(k) are the predictor coefficients and p is their number. The reflection coefficients, which are obtained as a byproduct of the solution, can be converted to log area ratios, then quantized and transmitted [4].

In warping the spectrum, it is practical (because of FFT algorithms) to compute the spectral values at equally spaced frequencies. This can be done by simple interpolation from the signal spectral values. The autocorrelation R(i) can then be computed via the FFT.

The procedure given above for linear predictive warping makes use of the autocorrelation method of linear prediction [5]. If the analysis is done using the covariance, lattice, or covariance lattice [6] methods, then the procedure has to be modified as follows: after solving for the predictor coefficients, compute the all-pole model spectrum, then continue the procedure starting at step (b) above. The all-pole model spectrum is given by:

$$P(\omega) = \frac{1}{\left|1 + \sum_{k=1}^{p} a(k) e^{-jk\omega}\right|^2}. \qquad (2)$$

3. Spectral Dewarping

At the receiver of the LPCW vocoder, the received parameters are decoded. If log area ratios are received, they are decoded into reflection coefficients, which are converted in turn to the corresponding predictor coefficients using a simple recursive procedure [5]. These coefficients correspond

to the warped spectrum and, therefore, cannot be used for synthesis. One must first perform the necessary dewarping.

The dewarping procedure is as follows:

(a) Using the decoded predictor coefficients, compute the all-pole model spectrum from (2).
(b) Dewarp this spectrum using the inverse of the function used in the original warping.
(c) Take the Fourier transform of the dewarped spectrum to obtain the corresponding autocorrelation function.
(d) Use this autocorrelation function in (1) to compute the predictor coefficients (and hence the reflection coefficients) corresponding to the dewarped spectrum. The number of poles (predictor coefficents) here can be as large as desired to approximate the dewarped spectrum.
(e) Synthesize the speech waveform using these computed coefficients.

After step (b) above it is possible to take a different route to obtain the parameters of the synthesis filter. Instead of using linear prediction, one could use the cepstrum [7] to compute the minimum phase impulse response whose spectrum is identical to the dewarped spectrum. This impulse response is then used for the synthesis filter.

Discussion

It is clear from the dewarping procedures given above that the amount of processing needed at the synthesizer is comparable to that of the analysis. This increase in computation relative to a regular LPC vocoder is certainly a disadvantage. Whether the extra expense is justified or not depends on the benefits achieved. For a given bit rate, the main benefit is an increase in the speech bandwidth representable using the same bit rate. This increase in bandwidth is on the order of 50%.

4. Warping Functions

Since linear predictive (LP) warping allows for arbitrary warping of the spectrum, one must choose a warping function appropriate for vocoder purposes. One reasonable warping function would transform the linear frequency scale to the mel scale [8], which compresses high frequencies relative to low frequencies.

The relation between the mel scale and frequency is shown in Fig. 1, which shows how subjective pitch (in mels) is related to frequency (in Hz) for pure tones up to 5 kHz. This relation is similar to those of critical band masking effects and equal intelligibility curves [8]. The mel-frequency relation can be approximated by the following equation

$$m = 2595 \log_{10}(1+\frac{f}{700}), \qquad (3)$$

where f is the frequency in Hz and m is the pitch in mels. The mel scale is adjusted such that m=1000 mels corresponds to f=1000 Hz.

Since, in our application, spectra are defined in the z plane, we need a warping function on the angle (which corresponds to frequency) in the z plane. This implies that we must assume a particular sampling frequency F. Let

$\omega = 2\pi\frac{f}{F}$ = original angle in the z plane corresponding to frequency f,

Ω = warped angle corresponding to f.

The warping function Ω is obtained from (3) by setting $\Omega = \pi$ for $\omega = \pi$ or $f=F/2$, half the sampling frequency. The result is:

$$\Omega = \pi \frac{\log_{10}(1+\frac{f}{700})}{\log_{10}(1+\frac{F}{1400})}, \quad 0\le f\le\frac{F}{2}. \qquad (4)$$

Note that the warping function in (4) is defined only up to f=F/2. For $F/2\le f\le F$, the function is taken to be the mirror image about the real axis. The mel warping function is plotted in Fig. 2 for a sampling frequency F=10 kHz, which corresponds to a speech bandwidth of 5 kHz.

The mel warping function could be used very profitably with a homomorphic vocoder [9] which employs cepstral warping or autocorrelation warping [10]. However, using the mel function with an LPCW vocoder seems to give unsatisfactory results. We believe the reason to be as follows. For LP to give best results, it is important that the all-pole model is well suited to the signal spectrum, which is true for a large and perceptually important class of speech spectra. If the signal spectrum is warped nonlinearly, then the all-pole model ceases to be a good spectral model. Therefore, the results are bound to be less than satisfactory. Note that this problem does not affect cepstral warping results, since cepstral warping is not based on a specific model.

The solution we offer to this problem in an LPCW vocoder is to have a warping function that is as linear as possible in the frequency range where the all-pole model is important, e.g. up to the third formant region. For higher frequencies the function can be quite nonlinear since only a rough estimate of the spectrum at those frequencies is needed. Fig. 2 shows a sine warping function

$$\Omega = \pi \sin(\frac{\pi f}{F}), \quad 0\le f\le\frac{F}{2}, \qquad (5)$$

which, for F=10 kHz, is nearly linear up to 2.5 kHz, and very nonlinear above that. One could, of course, design other warping functions that comprise more than a single curve.

5. Examples

Figs. 3 and 4 show two examples of using the sine warping function with spectra of the vowel [o] and the fricative [s], respectively. In each of the two examples, Fig. a is a 12-pole fit to the original spectrum, Fig. b is a 9-pole fit to the warped spectrum (shown after dewarping), and Fig. c is a 9-pole fit to the original spectrum. Note the greater

detail in the first formant region in Fig. 3b as compared to Fig. 3a, while the high frequency region is not matched as well in Fig. 3b. In comparing the two 9-pole fits, Figs. 3b and 3c, there is no doubt that Fig. 3b is a better "perceptual" fit to the spectrum, since the first three formants in Fig. 3b are better matched than in Fig. 3c. In contrast, Fig. 4c seems to be a better fit to the spectrum than 4b. But for a fricative, the match of Fig. 4b might be enough for good quality resynthesis.

These examples demonstrate that the use of spectral warping with an LPC vocoder could lead to a more efficient representation of the spectrum for the same speech quality. Although it might be practical to employ a fixed warping function for all situations, it is certainly possible to use several warping functions for different types of spectra. However, it is not clear that the possible increase in efficiency is worth the extra cost.

6. Conclusions

LPCW, an LPC vocoder with LP spectral warping, has been proposed. In this vocoder, a spectral warping function is used to compress high frequencies relative to low frequencies; a technique which is hypothesized to accommodate wider band speech signals. The result is improved speech quality for the same transmission rates.

Acknowledgment

This work was supported by the Information Processing Techniques Branch of the Advanced Research Projects Agency.

References

1. J.L. Flanagan, Speech Analysis Synthesis and Perception, Second Edition, New York: Springer-Verlag, 1972.

2. J. Roberts, C. Smith and R. Wiggins, "Triple-Function Voice Coder," J. Acoust. Soc. Am., Vol. 57, Supplement No. 1, S35, Spring 1975.

3. J. Makhoul, "Spectral Linear Prediction: Properties and Applications," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 283-296, June 1975.

4. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 309-321, June 1975.

5. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, Vol. 63, 561-580, April 1975.

6. J. Makhoul, "New Lattice Methods for Linear Prediction," IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, Pa., April 1976.

7. A. Oppenheim and R. Schafer, Digital Signal Processing, Ch. 10, New Jersey: Prentice-Hall, 1975.

8. J.C.R. Licklider, "Basic Correlates of the Auditory Stimulus," in Handbook of Experimental Psychology, (S.S. Stevens, ed.), 985-1039, New York: John Wiley and Sons, 1951.

9. A. Oppenheim, "A Speech Analysis-Synthesis System Based on Homomorphic Filtering," J. Acoust. Soc. Am., Vol. 45, 458-465, Feb. 1969.

10. J. Makhoul, "Methods for Nonlinear Spectral Distortion of Speech Signals," IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, Pa., April 1976.

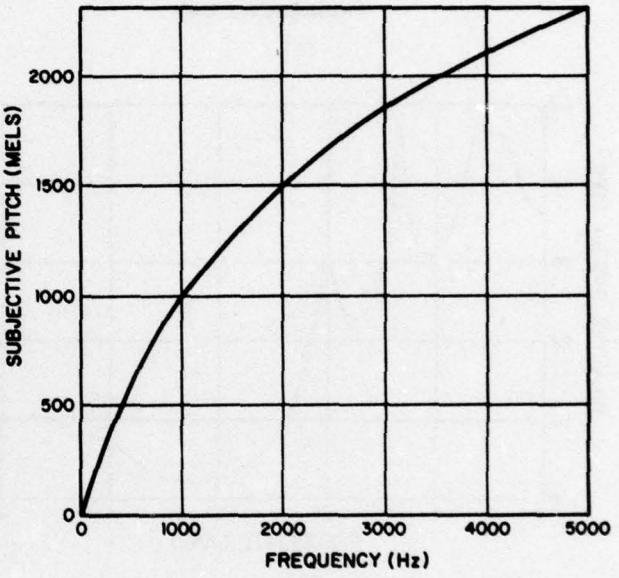


Fig. 1. Subjective pitch versus frequency.

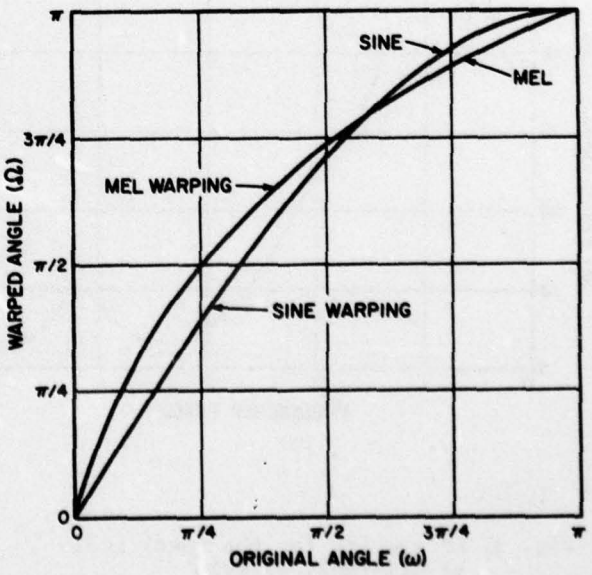
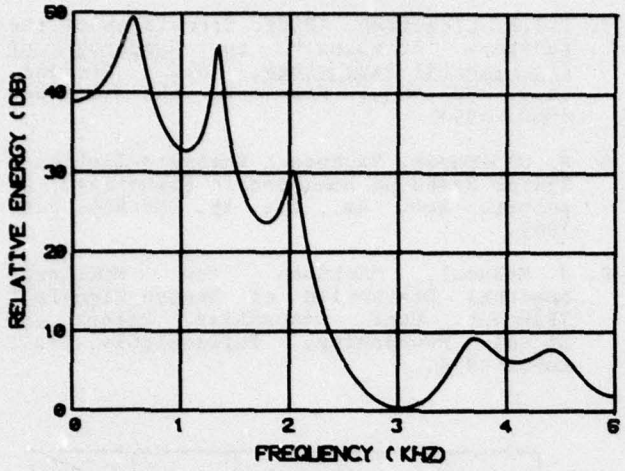
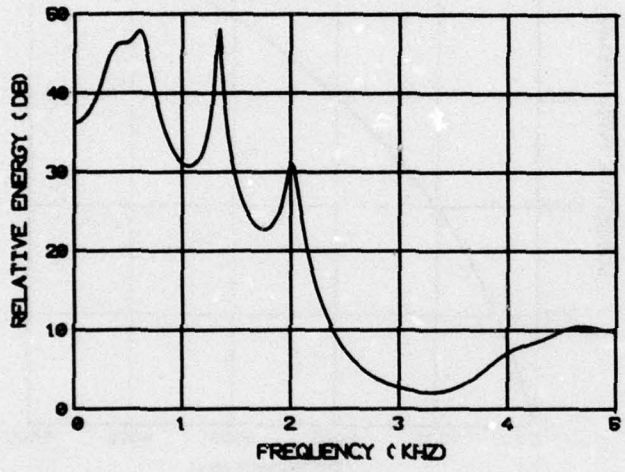


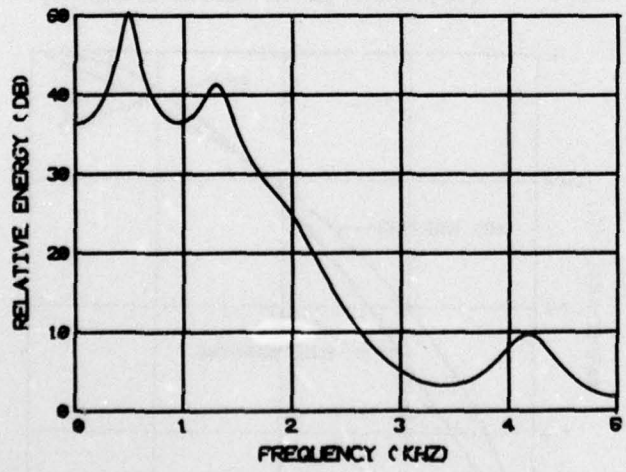
Fig. 2. Mel warping and sine warping.



(a)

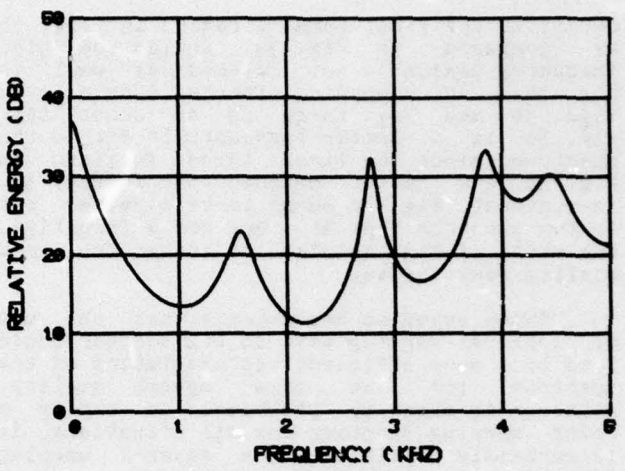


(b)

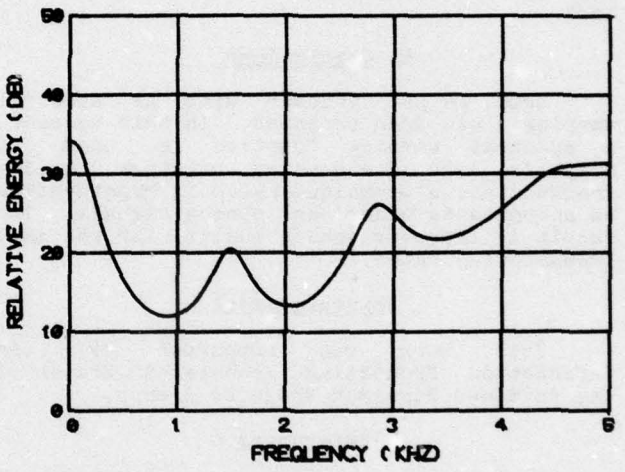


(c)

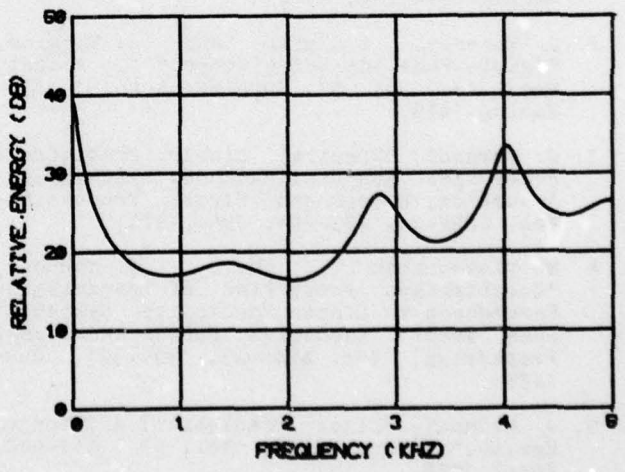
Fig. 3. LP spectra for the vowel [o].
a) Original, 12-pole
b) Warped, 9-pole (shown dewarped)
c) Original, 9-pole.



(a)



(b)



(c)

Fig. 4. LP spectra for the fricative [s].
a) Original, 12-pole
b) Warped, 9-pole (shown dewarped)
c) Original, 9-pole.

APPENDIX 6

THE APPLICATION OF A FUNCTIONAL PERCEPTUAL MODEL OF SPEECH TO VARIABLE-RATE LPC SYSTEMS

(Paper presented at the IEEE International Conference on on Acoustics, Speech, and Signal Processing, Hartford, CT, May 1977.)

THE APPLICATION OF A FUNCTIONAL PERCEPTUAL MODEL OF SPEECH TO VARIABLE-RATE LPC SYSTEMS

R. Viswanathan, J. Makhoul, and R. Wicke

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

A functional perceptual model is considered in which continuous speech is represented in terms of speech parameters extracted at a minimal set of time points or frames, not necessarily equally spaced, in such a way that the perceived quality of the resynthesized speech is no worse than that of the full, unreduced, parameter data from which the model parameter values are derived. The validity of this model has been experimentally demonstrated by the work of Olive and Spickenagel, using a phoneme-based, nonautomatic, method. In this paper, we describe the results of our work towards developing a fully automatic scheme for perceptual modeling of speech parametrically represented by LPC parameters. We present the model and the automatic scheme from the viewpoint of their application to efficient, variable frame rate, narrowband speech transmission.

1. Perceptual Model

In our work on developing minimally redundant narrowband speech transmission systems, we have used quite successfully the concept of variable frame rate (VFR) transmission [1,2,11,12]. In a VFR scheme, model parameters (LPC parameters, log pitch, log gain) are transmitted only when the properties of the speech signal have changed sufficiently since the preceding transmission; the parameters for the untransmitted frames are regenerated at the receiver through linear interpolation between the parameters of the two adjacent transmitted frames. For example, speech parameters may be transmitted less often during steady-state portions of speech, and more often during rapid speech transitions.

The concept of VFR transmission was applied to formant parameters of speech by McLarnon et al [3], and to LPC parameters of speech [4] by Magill [5] and by two of the present authors [1,2]. For VFR transmission of LPC parameters, the log likelihood ratio measure of Itakura [6] has been used for deciding which frames to transmit. In our work, linear predictive analysis was done once every 10 ms on speech, low-pass filtered at 5 kHz and sampled at 10 kHz, to extract 100 frames/sec (fps) of LPC data. Using the VFR scheme, we reduced the average frame rate of transmission of LPC data (excluding pitch and gain, which were transmitted at the full fixed rate of 100 fps) to about 37 fps, with only a small change in the quality of the resynthesized speech relative to the case when all the available 100 fps data were transmitted. Further, we observed that any significant reduction in the frame rate below 37 fps introduced, in general, noticeable distortions in the speech quality.

In an effort to reduce the average frame rate further, without speech quality degradation, we have recently based our work on VFR transmission on the following functional perceptual model of speech:

1) Speech can be represented in terms of LPC (or other) parameters extracted at a minimal set of perceptually significant time points (or frames), not necessarily equally spaced.

2) Between any two such time points, the parameters vary linearly.

3) The location of these points is obtained independently for pitch, gain, and spectral (or LPC) parameters.

Our requirement is that the quality of the resynthesized speech based on this model should be no worse than that of the unreduced or the full 100 fps case. The question then is: What is the minimal set of perceptually significant frames for LPC parameters that is consistent with our quality requirement? (While we have used an operational definition of minimal sets for pitch and gain in our work [11,12], we shall only discuss the LPC parameters in this paper.) The recent work of Olive and Spickenagel [7] suggests that the minimal set for LPC parameters is on the order of two frames per phoneme. (This corresponds to about 24 fps, assuming an average speech rate of 12 phonemes/sec.) In their work, they used a manual, trial-and-error scheme to locate the minimal

representative set, using LPC area parameters. Pair-wise comparison tests between the resynthesized speech at the resulting rate and that at the full 100 fps rate indicated no significant differences in perceived quality. Although the method described by Olive and Spickenagel was not automatic and involved trial and error adjustments, their work, nonetheless, provides an "existence proof" for what we call a perceptual model of speech, and supplies a reasonable lower limit to the average frame rate. (Very recently, Olive [9] reported on a semiautomatic method that employs an iterative frame elimination procedure over individual sentences, and which was found to yield about the same performance as did his manual approach [7].)

The goal of the work reported in this paper is to develop a fully automatic VFR scheme that (1) uses the information about the transmission parameters only, (2) results in an average frame rate that is close to the above-mentioned lower limit, and (3) produces speech whose quality is no worse than that of the speech synthesized using the full 100 fps LPC data. Towards achieving this goal, we first investigated a manual or nonautomatic scheme (Section 2), and then developed an automatic scheme based on the results and experience gained from the manual procedure (Section 3).

2. Manual Scheme

The main purpose of the manual scheme described below was to gain insights and ideas for developing transmission criteria for automatic perceptual modeling based on the information about the transmission parameters only. As transmission parameters, we used log pitch, log gain, and log area ratios (LARs). (For the many desirable properties of LARs, see [8].) In addition, we hoped that the results of our manual scheme would serve as another experimental validation of the perceptual modeling hypothesis.

As a key tool for manually carrying out the perceptual modeling task, we developed an interactive display program on our PDP-10/IMLAC PDS-1 computer facility. The program displays all the transmission parameters as well as the transmission status (0 or 1) of each of these parameters for every analysis frame, as functions of frame number. For any desired frame, the program can also display the values of displayed parameters, the power spectrum of the linear prediction filter, and the speech waveform in that frame. By viewing the displayed information for several utterances, one gains an intuitive feel for the magnitudes of parameter variation under various speech events and starts to develop simple rules that may be used in deciding whether or not a given frame of data should be transmitted. To further aid the user, we incorporated a number of features that allow the user to 1) manually mark selected frames of analyzed data for transmission, 2) synthesize speech from a specified amount of transmitted data, and 3) play out through a D/A converter specified portion of either synthesized speech or natural speech or both for on-line evaluation of relative speech quality.

Using the above interactive program, we accomplished the task of manually deriving the minimal set of frames for several utterances. Pitch, gain and 14 log area ratios, using the autocorrelation method [4] of linear prediction, were computed at a rate of 100 fps from speech sampled at 10 kHz. We selected a minimum number of frames of LAR data for transmission, out of the available 100 fps analysis data, using only the information about the transmission parameters and employing rules such as the following:

1) when log area ratios change roughly linearly, transmit them only for the frames corresponding to the endpoints of the line, since the LARs for the intermediate frames will be generated at the receiver through linear interpolation, and
2) ignore or deemphasize large changes in the values of LARs when the associated filter gain is low, since these low-gain frames have a relatively small effect on perception.

The overall objective was to reduce the frame rate as much as possible with the constraint that the resynthesized speech should be almost indistinguishable (as judged by informal listening tests) from the speech synthesized with all the analysis frames of data transmitted. We achieved a minimum frame rate of about 27 fps on the average, computed over 7 sentences of continuous speech from 4 speakers. In terms of phonemes, this rate came to about 2.2 frames/phoneme, which is slightly higher than the rate that Olive and Spickenagel reported [7].

Figure 1 shows the time plots of pitch in Hz (F0), speech signal energy per sample in dB (R0) and the first four LARs in dB (G1-G4), for the utterance "The trouble with swimming is that you can drown", spoken by a female. The long vertical lines mark the frames selected for transmission using the above manual approach.

3. An Automatic Scheme

After gaining confidence in our

manual approach, we developed an automatic scheme for selecting frames of LAR data for transmission, based on the results and experience gained from the manual scheme discussed above. As in the manual scheme, the automatic scheme uses only the information about the transmission parameters. An outline of the automatic scheme is presented below.

The automatic scheme employs a two-stage procedure for selecting frames for transmission. In the first stage, a chunk of successive analysis frames of data are considered; the number of frames in the chunk is variable, but its maximum can be specified. In the synthesis experiments discussed later, we chose a maximum of 9 frames. The decision to transmit a frame of data is made in the first stage as follows. Assume that frame n in the current chunk has been marked for transmission, and that frame (n+m) is under consideration. For each of the (m-1) frames that lie between frames n and (n+m), and considering the first LAR, G1, we compute the error between the actual value of G1 and the value obtained from linear interpolation between frames n and (n+m). These (m-1) errors are squared, weighted first by the speech signal energy (in dB) of the corresponding frame and then by a quantity which depends inversely upon the local rate of change of G1, and then finally averaged. This weighted average error is compared against a threshold. If the threshold is exceeded, frame (n+m-1) is marked for transmission; if not, the above procedure is repeated for G2, etc. The scheme considers up to G4 only; if the error does not exceed the threshold for all four LARs, it advances to frame (n+m+1) and the entire procedure is repeated. Of course, if a frame is marked for transmission, all the LARs are simultaneously transmitted.

The second stage of the automatic scheme considers the last transmitted frame in the previous chunk and those frames in the present chunk that have been marked for transmission, and attempts to eliminate any unnecessary transmissions. The decision procedure employed in the second stage is the same as in the first stage, except that now the time-averaged error is also averaged over the first four LARs. Our experiments indicated that the second stage deleted about 10% of the transmission marks decided by the first stage.

It should be pointed out that the choice of the various values of the weighting functions and the thresholds involved extensive experimentation; we optimized the choice by comparing against the transmission marks that were manually obtained, and by listening to the resulting synthesized speech.

We used the automatic scheme over the same speech utterances that we experimented with in our manual perceptual modeling approach. The average frame rate of LAR transmission with the automatic scheme came out to be about 26 fps. Although the average frame rates were approximately the same for the manual and automatic schemes, the actual locations of transmitted frames were, in general, different for the two cases. Informal listening tests conducted on the syntheses obtained from the manual and the automatic perceptual modeling approaches and from the fixed 100 fps system indicated that they all have roughly the same overall quality. An experienced listener could, for some utterances, pick the synthesis from the automatic scheme as being slightly inferior to the syntheses from the other two systems. We plan to modify some of the details of the automatic scheme to enhance the quality of the resynthesized speech.

4. Discussion

In all the synthesis experiments reported above, pitch and gain were transmitted at the full 100 fps rate and none of the transmission parameters were quantized. To investigate the perceptual model under parameter quantization and under conditions of narrowband speech transmission, speech was preemphasized and analyzed using an 11-th order LPC model; pitch and gain were quantized logarithmically using 6 bits and 5 bits, respectively; LARs were quantized using about 44 bits/frame [2]. Comparisons of syntheses from the fixed 100 fps system and the VFR system based on the perceptual model (manual or automatic scheme) indicated the following interesting, and perhaps surprising, result: For utterances for which LPC parameters vary relatively slowly in time (e.g., "Why were you away a year, Roy?"), the syntheses from the 100 fps system sounded worse, in particular had a more "wobble" quality, than those from the VFR system. We had had the same experience with our earlier VFR scheme that uses the log likelihood ratio criterion. Our explanation for the observed quality difference is that for slowly varying utterances, the error due to parameter quantization is more than the error due to interpolation. It is due to the above result that we required of the perceptually modeled speech to have a perceived quality that is <u>no worse</u> than that of the unreduced system from which it is derived, since we have seen that a lower rate VFR system can sometimes sound <u>better</u> than the unreduced system.

The perceptual model described in this paper deals with the problem of parametric representation of speech in time, which is perhaps the most important aspect of the overall problem of redundancy removal in speech as suggested by the results of a recent quality test [10]. We have investigated in detail several other aspects, which include variable order linear prediction [2,4], optimal parameter quantization and bit allocation [8], and Huffman coding of quantized parameters [2]. More recently, we have proposed VFR schemes for pitch and gain also [11]. When we employed all these compression techniques (without Huffman coding) with the automatic VFR scheme given in Section 3, we obtained good quality speech at average bit rates as low as 1700 bps, measured for continuous speech. With Huffman coding, we expect the average rate to drop below 1400 bps with absolutely no further reduction in quality.

Although we discussed above the VFR scheme as applied to efficient speech transmission, the scheme may be used in other applications such as speech storage and retrieval, speech synthesis by rule (as in the work of Olive and Spickenagel [7]), or for segmentation purposes in speech recognition.

Acknowledgments

The authors wish to thank R. Schwartz and W. Russell who implemented many of the features of the interactive display program described in Section 2. This work was sponsored by the Information Processing Techniques branch of the Advanced Research Projects Agency.

References

1. R. Viswanathan and J. Makhoul, "Towards a Minimally Redundant Linear Predictive Vocoder," Presented at the 88th Meeting of the Acoust. Soc. Amer., St. Louis, Nov. 1974.
2. J. Makhoul, R. Viswanathan, L. Cosell and W. Russell, Natural Communication with Computers, Final Report, Vol. II, Speech Compression Research at BBN, BBN Report No. 2976, Dec. 1974.
3. E. McLarnon, J.N. Holmes, and M.W. Judd, "Experiments with a Variable-Frame-Rate Coding Scheme Applied to Formant Synthesizer Control Signals," Preprints, Speech Commun. Seminar, Stockholm, Vol. 1, pp. 71-79, Aug. 1974.
4. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, Vol. 63, pp. 561-580, April 1975.
5. D.T. Magill, "Adaptive Speech Compression for Packet Communication Systems," Proc. Nat'l Telecommun. Conf., pp. 29D1-29D5, Nov. 1973.
6. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. ASSP, Vol. ASSP-23, pp. 67-72, Feb. 1975.
7. J.P. Olive and N. Spickenagel, "Speech Resynthesis from Phoneme-Related Parameters," J. Acoust. Soc. Amer., Vol. 59, pp. 993-996, April 1976.
8. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. ASSP, Vol. ASSP-23, pp. 309-321, June 1975.
9. J.P. Olive, "Semiautomatic Segmentation of Speech for Obtaining Synthesis Data," Presented at the 92nd Meeting of the Acoust. Soc. Amer., San Diego, Nov. 1976.
10. A.W.F. Huggins, R. Viswanathan and J. Makhoul, "Speech Quality Testing of Variable Frame Rate (VFR) Linear Predictive (LPC) Vocoders," Presented at the 92nd Meeting of th Acoust. Soc. Amer., San Diego, Nov. 1976.
11. R. Viswanathan, "Variable Frame Rate Transmission of Pitch and Gain," Appendix, BBN Report No. 3430, Sept. 1976.
12. E. Blackman, R. Viswanathan and J. Makhoul, "Variable-to-Fixed-Rate Conversion of Narrowband LPC Speech," to be presented at this conference.

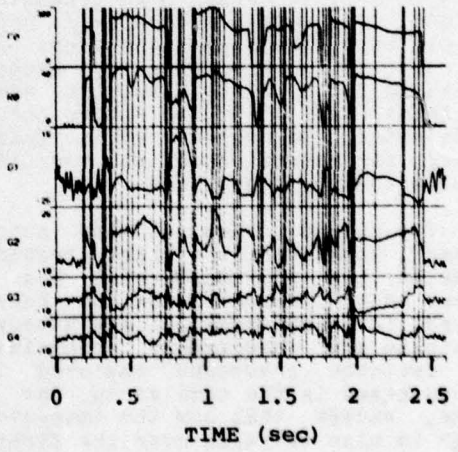


Figure 1. Time plots of transmission parameters for a sentence, along with transmission marks (long vertical lines) obtained by the manual scheme.

APPENDIX 7

A MIXED-SOURCE MODEL FOR SPEECH COMPRESSION AND SYNTHESIS

(Paper to be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK, April 1978.)

A MIXED-SOURCE MODEL FOR SPEECH COMPRESSION AND SYNTHESIS

J. Makhoul, R. Viswanathan, R. Schwartz and A.W.F. Huggins

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

ABSTRACT

This paper presents an excitation source model for speech compression and synthesis, which allows for a degree of voicing by mixing voiced (pulse) and unvoiced (noise) excitations in a frequency-selective manner. The mix is achieved by dividing the speech spectrum into two regions, with the pulse source exciting the low-frequency region and the noise source exciting the high-frequency region. A parameter F_c determines the degree of voicing by specifying the cut-off frequency between the voiced and unvoiced regions. For speech compression applications, F_c can be extracted automatically from the speech spectrum and transmitted. Experiments using the new model indicate its power in synthesizing natural sounding voiced fricatives, and in largely eliminating the "buzzy" quality of vocoded speech. A functional definition of buzziness and naturalness is given in terms of the model.

1. INTRODUCTION

Perhaps the single most important decision to be made in a pitch-excited speech compression system (vocoder) is the voiced/unvoiced (V/U) decision. Errors in this decision are readily perceived by the ear as a degradation of speech quality, and may also be accompanied by a loss in intelligibility. Yet, even if the V/U decision were somehow to be made "perfectly", the synthetic speech would continue to exhibit a distinct lack of naturalness, exemplified by a certain "buzziness" and a "lack of fullness." These characteristics are symptoms of the inadequacy of the binary V/U excitation model.

This paper explores the excitation problem in speech synthesis and presents a simple mixed-source model that allows for a degree of voicing. The new model is capable of producing more natural sounding speech; it seems to largely eliminate the buzziness problem and recover much of the fullness in the speech. In addition, it promises to reduce the adverse effects of voicing errors. A review of previous research relating to this model is given in a later section.

2. BASIC SYNTHESIS MODEL AND TERMINOLOGY

Throughout this paper, we shall assume the basic synthesis model shown in Fig. 1. In this model, a time-varying excitation signal excites a time-varying spectral shaping filter, the output of which is the synthetic speech. The excitation signal is assumed to have a flat spectrum, so that the spectral envelope of the synthetic speech is determined completely by the spectral shaping filter. Furthermore, we shall assume this model to hold for any type of synthesis, whether as part of a vocoder system or a synthesis system. In fact, we wish to argue below that our proposed source model is indeed adequate for both applications.

Restricting the excitation to have a flat spectrum necessarily limits us to two types of excitation: deterministic (pulse) or random (noise).

a) Pulse Source (Buzz)

The deterministic excitation is, in general, the impulse response of an all-pass filter, which we shall call an all-pass signal or pulse. The most trivial form of an all-pass pulse is a single impulse. When the pulse source produces a sequence of pulses separated by a pitch period, it is known as a buzz source. (Note that a single pulse could be used in the synthesis of the burst in a plosive sound [1]. However, the burst can also be synthesized *using the noise source*. We shall assume the latter in this paper; the pulse source will be used exclusively for buzz excitation.)

b) Noise Source (Hiss)

The random noise excitation may be the output of a random number generator. Generators with either a uniform or Gaussian probability distribution are readily available and are quite adequate. The noise source is also known as a hiss source.

Whether the actual excitation is buzz or hiss, or a combination of the two, one must always make sure that the excitation has a flat spectrum. We shall now describe how one might derive an appropriate source model by inspecting short-time speech spectra.

3. THE "IDEAL" SOURCE

For some particular speech signal, one can remove the short-time spectral envelope by appropriate inverse filtering, as shown in Fig. 2. The inverse filter A(z) can be obtained by cepstral techniques [2] or through the use of linear prediction [3]. The residual signal e(t) will then have a nominally flat spectrum. If in Fig. 1, the excitation u(t) is identical to the residual e(t), and the synthesis filter H(z) is the inverse of A(z), then the synthetic speech s'(t) will be identical to the original signal s(t).

However, for synthesis purposes, the synthetic signal need only sound like the original, and need not be identical to it. In addition, we need to manipulate the source pitch and to minimize the number of bits needed to represent the source. In order to accomplish this task, we first make use of an important property of speech perception, namely

CH1285-6/78/0000-0163$00.75@1978IEEE

that it is relatively insensitive to the short-time phase. Therefore, in order to model the residual e(t) to meet our requirements, we need only look at its spectrum and, except for pitch, disregard its phase structure for the moment.

Fig. 3 shows the signal power spectrum of 25.6 ms of a 10 kHz sampled signal in the middle of the vowel [I] in the word "list", and the corresponding residual spectrum. The residual was obtained by inverse filtering the speech signal with a 20th order linear prediction inverse filter. If somehow one could generate an excitation u(t) whose spectrum is identical to the residual spectrum, the synthetic speech would then sound (almost) the same as the original.

Therefore, our aim in developing source models will be to obtain an excitation spectrum that is as close as possible to the residual spectrum. Furthermore, we wish to obtain such an excitation spectrum using only the buzz and hiss sources described in Section 2. The source models will stem naturally from examining the characteristics of residual spectra.

4. CHARACTERISTICS OF RESIDUAL SPECTRA

In Fig. 3, the residual spectrum shows a clear periodicity up to about 3.5 kHz, and a lack of periodicity above that frequency. The periodicity corresponds to harmonics of the pitch fundamental frequency. By looking at residual spectra of other sounds it becomes amply clear that the existence of aperiodic frequency bands in sonorant sounds is quite common. While in Fig. 3 one can identify only two bands, it is possible to have several periodic and aperiodic adjacent bands in 5 kHz. For more examples, the reader is referred to the work of Fujimura [4], who studied voice aperiodicity by examining short-time signal spectra.

Partial devoicing of certain sounds is well-known from physical considerations. For example, the devoicing of [z] above about 1 kHz is well recognized and has long been taken advantage of in the synthesis of more natural voiced fricatives. On the other hand, it is also known that in the production of the tense front vowel [i], the constriction may become narrow enough to generate some turbulence, which is seen as devoicing of frequencies above about 3 kHz. However, most synthesizers to date have not taken advantage of this fact.

In addition to the foregoing types of sources of devoicing, Fujimura [4] has hypothesized that some of the spectral devoicing may be due to aperiodicities or irregularities in the vocal-cord movement. We have noticed that spectral devoicing often occurs during transitions between different sounds, including sonorant-sonorant transitions. In contrast to the examples given in the previous paragraph, we believe that the spectral devoicing due to vocal-cord irregularities and/or spectral transitions, may in fact be an artifact of the spectral estimation process. Whether such devoiced regions should be synthesized using a noise source is questionable.

In conclusion, residual spectra may be completely periodic (voiced), completely aperiodic (unvoiced), or may contain regions that are periodic and others that are aperiodic. The question now is how to model such spectra using the buzz and hiss sources.

5. PROPOSED SOURCE MODEL

One reasonable source model would divide the spectrum into a number of bands. Each band would then be excited by the buzz source if the band is considered periodic, and by the hiss source if the band is considered aperiodic. Fujimura [4] used a 3-band model in his experiment, and reported an improvement in speech naturalness. However, given our observations that spectral aperiodicities may not necessarily result from turbulent excitations, we have chosen a different model. In our model, we shall consider all spectral aperiodic regions that are in between two periodic regions to be in fact periodic. In other words, only the band above the periodic region with the highest frequency will be considered to be aperiodic and generated by a turbulent source. Our reasons for this choice are twofold: (a) Turbulent sources are more likely to excite higher frequencies; and (b) Excessive devoicing can be as degrading to quality as excessive voicing.

The resulting model is shown in Fig. 4. It is a mixed-source model with the buzz source exciting a time-varying low-frequency region of the spectrum, and the hiss source exciting the remaining high-frequency region. The selective excitations are realized by passing the pulse excitation through a low-pass filter with cutoff F_c, and the noise excitation through a high-pass filter with the same cutoff frequency F_c. The outputs of the two filters are then added, multiplied by the source gain and applied to the spectral shaping filter as the excitation signal. The model, then, has only two parameters: the cutoff frequency F_c, and the pitch period τ when $F_c > 0$. Since small changes in F_c are not perceptible, it is sufficient to quantize F_c into 2-3 bits for transmission purposes.

6. IMPLEMENTATION

a) Extraction of Source Parameters

The only difference between parameter extraction for the new source model and traditional pitch extraction is that the V/U binary decision has been replaced by the determination of a multi-valued parameter F_c in our model. The extraction of the pitch period is unchanged. Pitch period determination is relatively straightforward; many schemes exist that are quite adequate.

Just as V/U decision algorithms have proliferated, many algorithms will be developed that attempt to compute F_c in a perceptually satisfactory manner. The method we have chosen thus far is a peak-picking algorithm on the signal spectrum. The algorithm determines periodic regions of the spectrum by examining the separation between consecutive peaks and determining whether the separations are the same, within some tolerance level. F_c is taken to be the highest frequency at which the spectrum is considered to be periodic.

b) Filter Implementations

In our initial implementation we rounded the value of F_c to the nearest 500 Hz. Therefore, we needed lowpass and highpass filters with cutoff frequencies separated by 500 Hz. The filter designs were then stored and used in the synthesis as the need arose.

For each value of F_c, the 3 dB points for the lowpass and highpass filters were designed to be equal to F_c, in order that the spectrum of the

final excitation may be as flat as possible. The roll-off of the filters was considered to be of secondary importance, but should not be very sharp in any case. We considered FIR (finite impulse response) as well as recursive (low order Butterworth) filters. Both types of filters gave similar perceptual results.

7. RESULTS

Using the implementation described in Section 6, we compared the resulting syntheses to those using the binary V/U model in the context of a linear prediction (LPC) vocoder. A number of sentences from male and female speakers [5] were used in comparing the two analysis-synthesis systems. No quantization of parameters (except for F_c) was performed. One of the sentences had a concentration of fricative sounds "His vicious father has seizures," and another was a nonnasal sonorant sentence "Why were you away a year, Roy?" Other sentences were more general. With the V/U source, the fricative sentence sounded particularly buzzy for both male and female speakers, while the sonorant sentence was judged as buzzy only for low-pitched male speakers. The buzziness in both sentences was greatly reduced when using the mixed-source model. In general, the buzziness was always reduced with the new model. However, for some sentences the new synthesis produced certain small background noises. Upon careful listening, it was determined that some of those noises were present in the V/U synthesis but were masked by the buzziness. The other noises may be due to inaccurate determination of F_c and/or to the particular implementation of the model.

Overall, listeners thought that the new model performed better for female speakers (a pleasant surprise, for a change). The new synthesis was "raspier" and more in line with female speech which is considered to be more breathy than male speech.

A number of listeners reported that the new synthesis had a certain "fullness" that was absent with the V/U synthesis. We interpret this as an indication of the greater naturalness resulting from the new model.

8. REVIEW OF RELATED WORK

The only other work we know of where mixed excitation was used with LPC vocoders was that of Itakura and Saito [6]. But there, the two sources excited the whole spectrum simultaneously, with the "degree" of voicing being controlled by the relative amplitudes of the sources. The results were not encouraging [7].

After the development of our model over two years ago, we became aware of Fujimura's work [8,4], who as far as we know, was the first to suggest and test a frequency-selective mixed-source model. His work, which we mentioned earlier, was performed in the context of a pitch-excited channel vocoder. During the writing of this paper, Fujimura brought to our attention his other work with Kato et al. [9], where a variable cut-off frequency like ours was employed, but using a different algorithm to determine the cut-off. The work was done with a hybrid voice-excited and pitch-excited channel vocoder, and they reported excellent results. Coulter [10] used mixed excitation for the synthesis of voiced fricatives; however, the cut-off between the low and high frequency bands was fixed.

In speech synthesis, mixed excitation has been used routinely for the synthesis of voiced obstruents (see, for example, [1,11]). The parallel formant synthesizer of Holmes [1] allows for variable mixed excitation, and was especially used in transitions between unvoiced and voiced sounds. Upon careful reading, it became clear to us that the spirit of Holmes' synthesizer is similar to ours, except that the controls in his case are more complicated. A more recent hardware synthesizer by Strube [12] allows for mixed excitation using a single variable RC-circuit.

There have been numerous attempts at reducing buzziness by changing the shape of the pulse in voiced excitation, but to no avail. Recently, Sambur et al. [13] reported a reduction in buzziness by changing the pulse width to be proportional to the pitch period. Unfortunately, changing the pulse width changes the excitation spectrum; the effect is that of a variable lowpass filter. Spectrally flattening the pulse before excitation cancelled the reduction in buzziness [14].

9. DISCUSSION

a) Buzziness and Naturalness

It is interesting that the mixed-source model appears to reduce two seemingly different types of buzziness: the buzziness in voiced fricative synthesis, and the buzziness in sonorant synthesis associated mainly with low-pitched voices. Our hypothesis is that the two types of buzziness, in fact, result from the same process: that of an excess in buzz source excitation. Thus, our general rule is that:

too much buzz ⟶ "buzziness"
too much hiss ⟶ "breathiness" or "raspiness"

where the arrow is to be read as "results in". If more of the spectrum is excited by the buzz source than is necessary for naturalness, the result is buzziness. Similarly, if there is more hiss excitation than is necessary for naturalness, the result is breathiness or raspiness. This leads us to a functional definition of naturalness, as it relates to mixed excitation:

> Naturalness is achieved by that proper mix of buzz and hiss excitations that leads to a synthesis that is neither buzzy nor breathy or raspy.

b) Modulation and Naturalness

Certain synthesizers, such as that of Klatt [11], modulate the hiss source by the buzz source for the synthesis of voiced fricatives. While it is known that the noise source in the vocal tract is in fact modulated by the vocal cord output, it is not clear that such modulation is necessary for achieving naturalness in synthetic speech. Whatever effect modulation has, it appears to be of a secondary nature. The synthesizer of Holmes [1] does not contain any modulation, and he reported very natural speech synthesis. Although initially we included modulation in our model, it is our opinion at this point that source modulation is not necessary for natural synthesis, and therefore we have decided not to incorporate it as part of the model.

c) Phase and Naturalness

It is generally agreed that proper phase determination of buzz excitation should lead to more natural synthesis. Furthermore, such phase cannot be in the form of some "optimal" pitch pulse shape. The phase must change from one pitch pulse to the next in some appropriate manner. Thus far, our model calls for an all-pass pulse, but does not specify the phase. Exactly how the phase should change between pulses is a subject for future research.

10. CONCLUSION

We have presented a frequency-selective mixed-source excitation model for use in both speech compression and speech synthesis. The model has a single continuous parameter, F_c, which divides the spectrum into two regions, with the buzz source exciting the low frequency region below F_c, and the hiss source exciting the high frequency region above F_c. Naturalness (no buzziness or breathiness) is achieved by the proper mix of the two sources, i.e., by the proper determination of F_c.

ACKNOWLEDGMENTS

The authors wish to thank K.N. Stevens for many discussions, especially during the initial development of the model. One of the authors (JM) had a useful discussion with L. Rabiner concerning the different types of buzziness in vocoded speech. This work was sponsored by the Information Processing Techniques branch of the Advanced Research Projects Agency under Contract No. MDA903-75-C-0180.

REFERENCES

1. J.N. Holmes, "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer," IEEE Trans. Audio and Electroacoust., pp. 298-305, June 1973.
2. A.V. Oppenheim and R.W. Schafer, Digital Signal Processing, Prentice-Hall Inc., 1975.
3. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, pp. 561-580, April 1975.
4. O. Fujimura, "An Approximation to Voice Aperiodicity," IEEE Trans. Audio and Electroacoust., pp. 68-72, March 1968.
5. A.W.F. Huggins, R. Viswanathan and J. Makhoul, "Speech-quality testing of some variable-frame-rate (VFR) linear-predictive (LPC) vocoders," J. Acoust. Soc. Am., pp. 430-434, Aug. 1977.
6. F. Itakura and S. Saito, "Analysis Synthesis Telephony Based upon the Maximum Likelihood Method," Reports of 6th Int. Cong. Acoust., Tokyo, Japan, Paper C-5-5, pp. C17-20, 1968.
7. F. Itakura, personal communication.
8. O. Fujimura, "Speech Coding and the Excitation Signal," 1966 IEEE Int. Comm. Conf., Digest of Technical Papers, p. 49, 1966.
9. Y. Kato, K. Ochiai, O. Fujimura and S. Maeda, "A Vocoder Excitation with Dynamically Controlled Voicedness," 1967 Conf. Speech Comm. and Processing, Cambridge, Mass., pp. 288-291, 1967.
10. D. Coulter, Application of Simultaneous Voice/Unvoice Excitation in a Channel Vocoder, U.S. Patent No. 3903366, 1975.
11. D.H. Klatt, "Structure of a Phonological Rule Component for a Synthesis-by-Rule Program," IEEE Trans. Acoustics, Speech and Signal Processing, pp. 391-398, Oct. 1976.
12. H.W. Strube, "Synthesis Part of a 'Log Area Ratio' Vocoder in Analog Hardware," IEEE Trans. Acoustics, Speech and Signal Processing, pp. 387-391, Oct. 1977.
13. M. Sambur, A. Rosenberg, L. Rabiner and C. McGonegal, "On Reducing the Buzz in LPC Synthesis," 1977 IEEE Int. Conf. Acoustics, Speech and Signal Processing, Hartford, Conn., pp. 401-404, 1977.
14. B. Atal, personal communication.

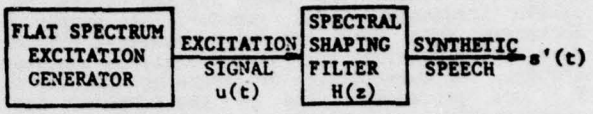


Fig. 1 Basic synthesis model.

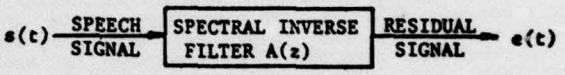


Fig. 2 Inverse filtering the speech signal to obtain a residual signal with a flat spectrum.

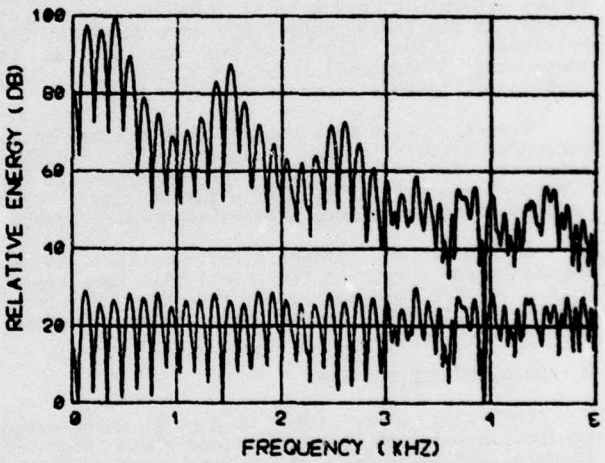


Fig. 3 Signal spectrum (top) and residual spectrum (bottom) for the vowel [I] in the word "list".

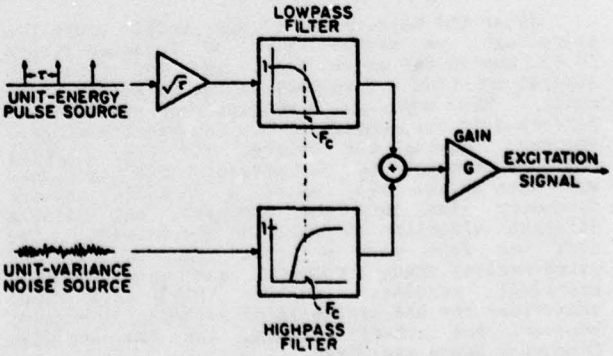


Fig. 4 Frequency-selective mixed-source excitation model.

APPENDIX 8

EXTENDED SET OF PHONEME-SPECIFIC TEST SENTENCES

Phoneme-Specific Sentences

Key

All voiced, sonorant
- Type 11: glides = w, r, y + vowels (no zeroes)
- Type 12: glides = w, r, y, + l

All voiced, sonorant, with zeroes
- Type 21: nasal = m, n, ng
- Type 22: nasals + l = m, n, ng, l
- Type 23: glides + nasals = w, r, y, l, m, n, ng

Stops and Affricates
- Type 41: voiced stops = b, d, g
- Type 42: voiced stops + affricate = b, d, g, j
- Type 43: unvoiced stops = p, t, k
- Type 44: unvoiced stops + affricate = p, t, k, ch
- Type 45: stops + affricates = b, d, g, j, p, t, k, ch.

Fricatives
- Type 51: voiced fricatives = v, dh, z, zh
- Type 52: unvoiced fricatives = f, th, s, sh
- Type 53: fricatives = f, th, s, sh, v, dh, z, zh

Place
- Type 61: all labials = p, b, f, v, w, m
- Type 62: all tongue-tip = t,d,th,s,sh,dh,z,zh,n,ch,j,l,r,y

Voicing
- Type 71: voiced stops, affric, frics = b, d, g, j, v, dh, z, zh
- Type 72: unvoiced stops,affric,frics = p, t, k, ch, f, th, s, sh
- Type 75: all voiced = b,d,g, j, v,dh,z,zh, m,n,ng, l,r,w,y
- Type 76: all unvoiced = p,t,k, ch, f,th,s,sh, h

All
- Type 80: Miller & Nicely Demos = all consonants except ch, j, h, y
- Type 81: All consonants

Sentences

Type 11: glides = w, r, y + vowels

* 1. Why were you away a year, Roy?
2. Why were you weary?

Type 12: glides = w, r, y, + l

1. Why were you all weary?
2. Our lawyer will allow your rule.
3. Our rule will allow you a lawyer.
4. We really will allow you a ruler.

Type 21: nasal = m, n, ng

* 1. Nanny may know my meaning.
2. Many young men owe money.
3. I'm one man among many.
4. When may we know your name?
5. I'm naming my own mine.
6. No-one knowing my name.
7. I'm no mean man.
8. I know many a mean man.
9. I know many mean men.
10. One name among many.
11. I'm known among men.
12. A man on a moon.
13. I know no minimum.
14. I'm naming one man among many.
15. I'm owing no-one any money.
16. Anne 'n May own many.
17. Anne 'n Arnie own one. (n only)

Type 22: nasals + l = m, n, ng, l

1. I'm well known among men.
2. Nine men moaning all morning.

Type 23: glides + nasals = w, r, y, l, m, n, ng

1. Where were we all wrong?
2. You were wrong all along.
3. I'll warn Ron away.
4. I know you're a loner.
5. I know you're all alone.
6. I really mean weighing in.
7. Why are you naming Wally?
8. When will our yellow lion roar?
9. Will you ring an alarm?
10. A morning alarm rang.
11. An alarm rang in only one room.
12. An alarming rule.
13. We may allow a new ruling.
14. A lawyer may well allow a new ruling.
15. An alarm rang a warning.
16. You will alarm me no more!
17. I'm learning a (my) new role.
18. I know you're really alone.
19. I'll remain in my narrow room.
20. We'll rely on no-one.
21. Anyone may rely on a mail-man.
22. I'll wear a maroon ring.
23. I'm wearing my maroon ring.
24. We'll remain all morning.
25. You'll remain in your room all morning.
26. We're all in mourning.
27. We'll allow you a new loan.
28. You're learning a new rule.
29. I'll lie in an alarming manner.
30. Why lie, when you know I'm your lawyer?
31. Any animal may run away.
32. A normal animal will run away
33. Mail me an aluminum railing.
34. Marilyn alone will marry me.
35. I'll willingly marry Marilyn.
36. I'm more normal in early morning.

Type 41: voiced stops = b, d, g

1. Do you abide by your bid?
2. Grab a doggie bag.
3. A greedy boy died.
4. Dad would buy a big dog.
5. Bobby did a good deed.
6. I begged Dad buy a dog.
7. Why did Gay buy a bad egg?
8. Did Bobby do a good deed?
9. Buy Dad a bad egg.

Type 42: voiced stops + affricate = b, d, g, j

1. Did George do a good job?
2. Greg adjudged Bobby dead.

Type 43: unvoiced stops = p, t, k

1. Kate typed a paper.
2. Take a copy to Pete.
3. Pat talked to Kitty.
4. Quite a cute act.
5. Peter took out a potato.
6. Patty cut up a potato cake.

Type 44: unvoiced stops + affricate = p, t, k, ch

1. Chip took a picture.
2. Teacher patched it up.
3. Chat quietly to teacher.
4. Quite quiet at church.
5. Catch a paper cup.
6. Actuate a paper copier.
7. Teacher taped up a packet.
8. Keep quite a cute picture.
9. Keep quiet at church.
10. Capture a cute puppy.
11. Teacher typed up a paper.
12. Katie tacked up a cute picture.

Type 45: stops + affricates = b, d, g, j, p, t, k, ch.

* 1. Which tea-party did <u>Baker</u> go to?
2. We'd better buy a bigger dog.
3. Georgie had to chew tobacco.

Type 51: voiced fricatives = v, dh, z, zh

1. View these azure vases.
2. They use our azure vials.
3. There's our azure vial.
4. There's usually a valve.
5. Those waves veer over.

Type 52: unvoiced fricatives = f, th, s, sh

1. I saw three fish.
2. A thief saw a fish.
3. Three chefs face a thief.

Type 53: fricatives = f, th, s, sh, v, dh, z, zh

* 1. His vicious father has seizures.
2. Whose shaver has three fuses?
3. Three of the chefs saw the thieves.

Type 61: all labials= p, b, f, v, w, m
(none)

Type 62: all tongue-tip = t,d, th,s,sh, dh,z,zh, n,

1. The judge's harsh decision really touched the youth.
2. Each decision shows the jury he lies through his yellow teeth.
3. Such a rash allusion to dosage teases the youth.
4. Seth yawns at each rash allusion to tne dosage.
5. The designers really earned the judge's derision this year.
6. Each allusion to Daisy's agility lessens her attention.
7. Each decision shows that he lies through his yellow stained teeth.
8. John drowned his sorrows in gin and orange juice.

Type 71: voiced stops, affric, frics = b, d, g, j, v, dh, z, zh
(none)

Type 72: unvoiced stops,affric,frics = p, t, k, ch, f, th, s, sh
(none)

Type 75: all voiced = b,d,g, j, v,dh,z,zh, m,n,ng, l,r,w,y

1. Does John believe you were measuring the gun?
2. Your brother's vision was gradually dimming.
3. The regular division was led by a young major.
4. I gather you will be abandoning the major revisions?
5. The young major's evasions were growing bolder.

Type 76: all unvoiced = p, t, k, ch, f, th, s, sh, h

1. I hope she chased her fox to earth.
2. A thickset officer pitched out her hash.
3. He checked through fifty ships.
4. She swiftly passed a health check.
5. He steps off a path to cash a check.

Type 80: Elliptic sentences (all except h,ch,j,y), from Miller, G. A. & Nicely, P. A., An analysis of perceptual confusions among some English consonants, J. Acoust. Soc. Amer. 1955, 27, 338-352.

1. The wealthy banker from Persia should be a good citizen.
2. The issue of McCarthy is forcing a great division among Republicans.
3. Division can be a fast operation with logarithms.
4. She thinks she bought some good rouge and lipstick from one of these men.

Type 81: All consonants.

1. If the treasure vans got so much publicity, we think you should hide your share.
2. The voyagers have ground the crankshaft with (th) unimpeachable precision.
3. The old-fashioned jacket was giving you both so much humorous pleasure.
4. Disillusioned taxpayers think the average gambler half wishes to cheat.
5. The average disillusioned gambler thinks he wishes for a cheap yacht.
6. Nothing could be further from reality than his illusion of chasing your gorgeous sheep away.
7. She thinks even the pale rouge you bought was much too gaudy for her age.

APPENDIX 9

QUALITY RATINGS OF LPC VOCODERS: EFFECTS OF NUMBER OF POLES, QUANTIZATION, AND FRAME RATE

(Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, CT, May 1977.)

QUALITY RATINGS OF LPC VOCODERS: EFFECTS OF NUMBER OF POLES, QUANTIZATION, AND FRAME RATE

A.W.F. Huggins, R. Viswanathan, and J. Makhoul

Bolt Beranek and Newman Inc.
50 Moulton Street, Cambridge, Mass. 02138

Four values for number of poles (13, 11, 9, 8) were combined factorially with three values of step size for quantization of log area ratios (0.5, 1, 2 dB), and with four values of frame rate (100, 67, 50, 33 per second), to define 48 LPC vocoder systems with overall bit rates ranging from 8.7 down to 1.3 kbps. Subjects rated the DEGRADATION of signal quality by each vocoder, for each of seven sentence tokens, chosen to challenge LPC vocoders maximally. The results define the combination of LPC parameters yielding the best speech quality for any desired overall bit rate.

1. Introduction

This study was performed to measure how the quality of LPC vocoded speech is affected by three different methods of reducing bit rate. These were:

1) reducing the number of poles used for spectral matching,
2) coarsening the step size used in quantizing the coefficients (log area ratios, Viswanathan & Makhoul, 1975),
3) reducing the number of frames of coefficients transmitted per second.

To establish the best operating point, for a range of different bit rates, it is necessary to perform a factorial study, in which each value of a parameter occurs with every combination of values of the other parameters. We used the following set of parameter values: Number of Poles, P: 13, 11, 9, or 8; Quantization Step Size, Q: 0.5, 1.0, or 2.0 dB; and Frame Rate, R: 100, 67, 50, or 33 per second, yielding 48 LPC systems (4x3x4). Two additional systems were included. One was an LPC system with 13 poles, quantization step size of 0.25 dB, and transmission rate of 100 frames per second. The other consisted of PCM speech at 110 kbps (i.e. the waveform sampled at 10 kHz and quantized to 11 bits), to act as an undegraded anchor. The bits per frame for each combination of number of poles and quantization step size appear in Table 1.

Quantization Step Size	No. of Poles 13	11	9	8
0.25 dB	76	--	--	--
0.5 dB	63	55	47	43
1.0 dB	50	44	38	35
2.0 dB	37	33	29	27

Table 1: Bits per frame for all combinations of number of poles and quantization step size used in the present study (excluding pitch and gain).

Pitch and gain were transmitted at the same frame rate as the coefficients. The overall bit rate for any system is calculated by adding 6 bits of pitch coding and 5 bits of gain to the bits per frame, and multiplying by the appropriate frame rate. The overall bit rate of the LPC systems ranged from 8700 bps (P = 13, Q = 0.25 dB, R = 100/sec), down to 1267 bps (P = 8, Q = 2.0 dB, R = 33/sec). Note that these rates do not include the benefits of Huffman coding, in which the most frequently used values are assigned the shortest codes. This procedure can further reduce bit rates by about 20%, with absolutely no change in the coefficient values transmitted (Makhoul et al, 1974).

2. Sentence Materials

Our earlier subjective quality tests showed the necessity of passing all sentence materials through all systems (Huggins & Nickerson, 1975). Other researchers have reached similar conclusions (Pachl et al, 1971). In our earlier tests, we developed a set of six sentences, each read by six talkers, that was both representative, in that it covered a wide range of speech events and talker characteristics, and also challenging, in that some speech material was included that would fully extend any LPC vocoder's abilities. Unfortunately, we could not use all 36 speaker-sentence combinations in the present study, since passing them through all 50 vocoder systems would have made the study

unmanageably large. We therefore selected a subset of seven speaker-sentence combinations, and confirmed that they were adequately representative of the full set by repeating the MDPREF analysis using just the data from the subset.

The subset of sentence tokens that was selected consisted of: JB1, DD2, RS3, AR4, JB5, DK6, and RS6, where the initials identify the speaker and the number identifies the sentence. Relevant details of the sentences, and of the speakers' voices, are given in Table 2.

ID	F0	Sentence
JB1	119	Why were you away a year, Roy?
DD2	134	Nanny may know my meaning.
RS3	195	His vicious father has seizures.
AR4	165	Which tea-party did Baker go to?
JB5	124	The little blankets lay around on the floor.
DK6	97	The trouble with swimming
RS6	193	is that you can drown.

Table 2: The seven stimulus sentences, with the speaker's average fundamental frequency in Hz.

3. Generation of Stimulus Tapes

Each of the seven input sentences was digitized (11 bits, 10 kHz), and passed through each of the 50 simulated vocoder systems, to yield a total of 350 different stimulus items.

Earlier studies have demonstrated that a subject's judgment, especially of speech stimuli, can be strongly affected by the preceding stimulus (e.g. Huggins, 1968). It is important to control for effects such as this by counterbalancing the presentation order. A complete counterbalancing of the 50 vocoder systems was generated, in which every system followed every other system once, with independent approximate counterbalancing of the sentences. This required only seven passes through the 350 stimuli, and had the further advantage that even within each pass, all ranges of contrast between successive stimuli occurred equally often, so that no severe departures from balance occurred even within one pass. The sequence was generated by a trial and error search, following an algorithm described by Williams (1950). No system and no sentence followed itself.

We tried to further reduce sequence effects (and thus improve the reliability of the data) by a novel method. A continuous speech babble, at the same level as the speech, was automatically faded in and out again during the inter-stimulus interval. We hoped that, by analogy with the "suffix" effect found in studies of auditory short term memory (Crowder & Morton, 1969), the babble would interfere with the memory trace of earlier stimuli, on which sequence effects presumably depend. The babble was developed at BBN for other purposes (Kalikow et al, 1976). The babble signal was recorded on a separate track of the tape, to permit the signal to be played with or without the babble.

Seven experimental tapes were then recorded. Stimuli were presented in blocks of ten, at a rate of one every 7.5 seconds, with a longer gap between blocks.

4. Experimental Procedures

The subject's task was to rate the degradation of the stimuli he heard. This negative attribute was chosen for scaling, as in our earlier experiment, because the scale has a natural origin, or zero, corresponding to undegraded speech. Instead of assigning a number to his judgment, the subject made his response by making a mark on a 10 cm line on his answer sheet. Two visual anchors were provided on the response line. The left anchor was 4 mm from the left end of the line, and was marked "PERFECT". The right anchor was 1 cm from the right end of the line. For data analysis, the response was converted into the distance in millimeters from the left end of the line (not the anchor) to the subject's mark where it crossed the response line. Thus small numbers correspond to high quality, and large numbers to poor quality.

Nine subjects served in the experiment. They were recruited by local university summer placement offices: all reported having normal hearing. Three of the subjects made the first five passes through the 350 stimuli, and six more subjects made only the first two passes.

5. Results

First, to check on the reliability of the data, the responses collected on each pair of passes through the 350 stimuli were correlated, for each subject. All correlations were significant well beyond $P<.001$, with the (Pearson product-moment) coefficients lying between 0.48 and 0.83, almost all of them in the top half of this range.

The mean degradation rating was calculated for each system, and these are plotted as a function of overall bit rate in Figs 1, 2, and 3. Each system is identified by three digits, corresponding

to the parameter level for P, Q, and R, respectively. Thus system 231 used level 2 of P (11), level 3 of Q (1.0 dB) and level 1 of R (100/sec), as shown in the key to the figure. The means of the ratings (N.B. not the ratings) have standard deviations of about 1.5 points. Therefore any difference between two plotted means that is larger than about 3-4 points is probably significant at P<.05 by t-test. (In fact it is likely that much smaller differences are also significant, since this test does not partial out the variability due to the sentence and subject. This can be done by comparing ratings on the pair of systems of interest before pooling across subjects and sentences.)

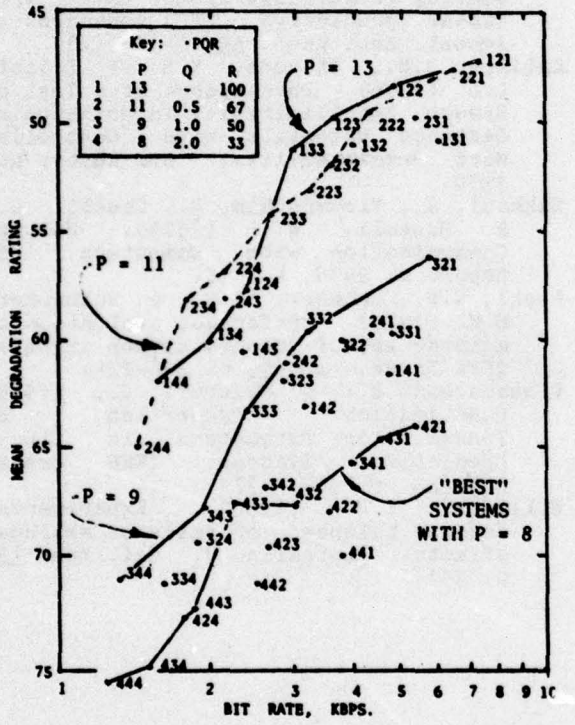


Figure 1: Mean degradation rating vs. Bit Rate for 48 LPC vocoders. Lines join "best" systems for each No of Poles.

In Figure 1, a line joins the "best" systems using 13 poles, and other lines join the best systems using 11, 9, and 8 poles. From inspection of Figure 1, it is clear that 13-pole systems give (slightly) better quality than 11-pole systems for most bit rates above 2750. 11-pole systems are (slightly) superior between about 1500 bps and 2750 bps. These differences are small, however, and are probably not significant. The best 11 and 13 pole systems are substantially better than the best 8 or 9 pole systems at comparable bit rates. These differences are large and highly reliable. The reason is that there is a highly significant interaction between the sex of the talker (or the talker's fundamental frequency) and the number of poles. This confirms earlier findings (Huggins & Nickerson, 1975; Huggins et al, 1976). Averaging ratings across all systems with the same number of poles shows that reducing the number of poles from 13 to 8 had relatively little effect on quality, for the three sentences spoken by females (RS3, AR4, RS6), whereas there is a massive reduction of quality for male voices when the number of poles is reduced below 11.

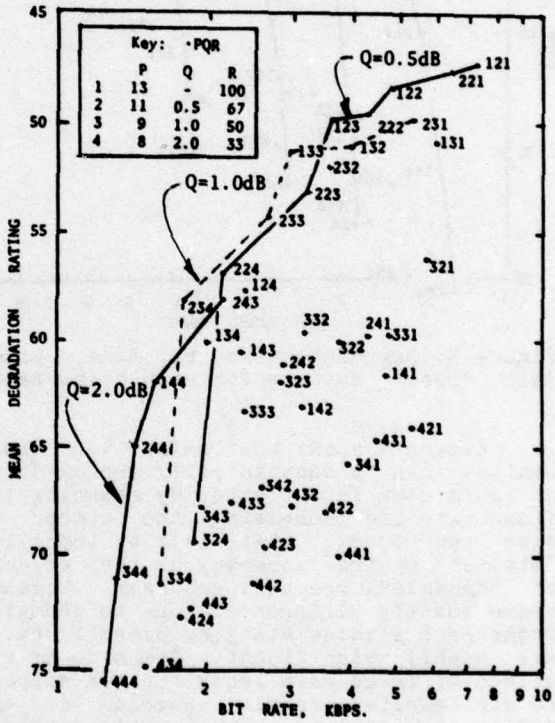


Figure 2: Degradation vs. Bit Rate. Lines join "best" systems for each Quantization Step Size.

Figures 2 and 3 present comparable plots, with best systems joined for each level of quantization, and for each level of frame rate, respectively. The differences in quality between different levels of quantization, at a given bit rate, are significant only at the very low bit rates. Here, quality is less affected

by coarsening quantization than by reducing the number of poles.

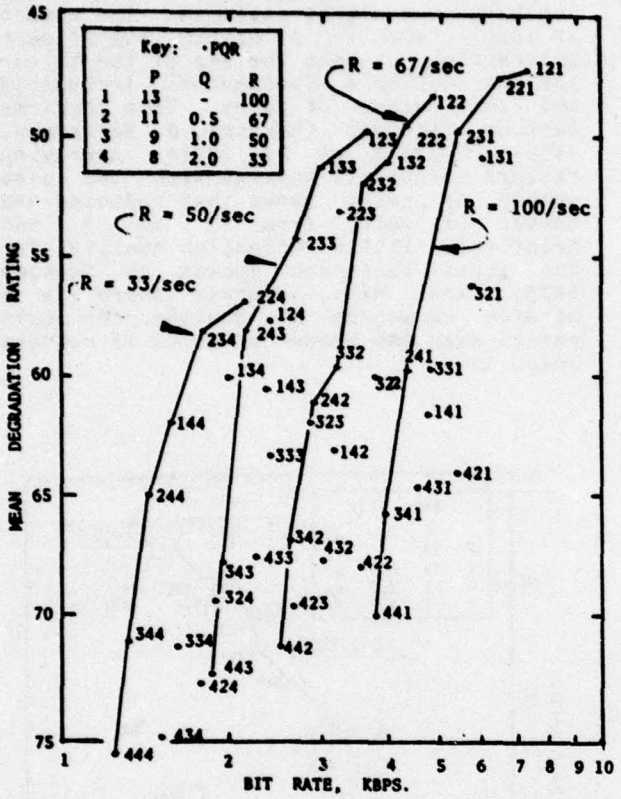


Figure 3: Degradation vs. Bit Rate. Lines join "best" systems for each Frame Rate.

Figure 3 shows that below 4.5 kbps, quality can be substantially improved, at no extra cost in bit rate, by reducing the frame rate and increasing the number of bits per frame, that is, by improving "static" spectral accuracy at the expense of "dynamic" spectral accuracy. Most of these quality differences, due to changing frame rate without changing overall rate, are highly significant. The size of the effect of frame rate lends further support to our earlier result (Huggins et al, 1976), suggesting that a well designed variable frame rate transmission scheme should yield substantial savings in bit rate without appreciable loss of quality.

Further analyses of these data, including multidimensional analysis, will be reported in a separate paper.

Acknowledgment

It is a pleasure to thank Paul Weene for his substantial help in performing this experiment. The research was sponsored by the Information Processing Techniques branch of the Advanced Research Projects Agency.

6. References

Crowder, R.A & Morton, J. (1969). Precategorical Acoustic Storage. Perception & Psychophysics 5, p. 365-375.

Huggins, A.W.F. (1968). The Perception of Timing in Natural Speech I: Compensation within the Syllable. Language & Speech, 11, p. 1-11.

Huggins, A.W.F., & Nickerson, R.S. (1975). Some effects of speech materials on vocoder quality evaluations. J. Acoust. Soc. Amer. 58, S-129 (A).

Huggins, A.W.F., Viswanathan, R., & Makhoul, J. (1976). Speech quality testing of variable frame rate (VFR) linear predictive (LPC) vocoders. J. Acoust. Soc. Amer. 60, S-108 (A).

Kalikow, D.N., Stevens, K.N. & Elliot, L.L. (1976). Development of a Test of Speech Intelligibility in Noise Using Sentence Materials with Controlled Word Predictability. BBN Report No. 3370.

Makhoul, J., Viswanathan, R., Cosell, L., & Russell, W. (1974). Natural Communication with Computers, BBN Report No 2976, Vol II.

Pachl, W.P., Urbanek, G.E. & Rothauser, E.H. (1971). Preference evaluation of a large set of vocoded speech signals, IEEE Trans. AU-19, p. 216-224.

Viswanathan, R., & Makhoul, J., (1975) Quantization Properties of Transmission Parameters in Linear Predictive Systems. IEEE Trans. ASSP-23, p. 309 - 321.

Williams, E.J. (1950). Experimental designs balanced for pairs of residual effects. Australian J. Sci. Res. A3, p. 351.

APPENDIX 10

SPEECH-QUALITY TESTING OF SOME VARIABLE-FRAME-RATE (VFR) LINEAR-PREDICTIVE (LPC) VOCODERS

(Paper published in the Journal of the Acoustical Society of America, Vol. 62, August 1977.)

Speech-quality testing of some variable-frame-rate (VFR) linear-predictive (LPC) vocoders[a)]

A. W. F. Huggins, R. Viswanathan, and J. Makhoul

Bolt Beranek and Newman Incorporated, 50 Moulton Street, Cambridge, Massachusetts 02138
(Received 22 February 1977; revised 25 April 1977)

VFR transmission of LPC vocoder coefficients is a technique developed to reduce the average transmission rate without appreciable loss of quality. The technique transmits parameters at a variable rate in accordance with the changing characteristics of the speech signal. In order to assess the effectiveness of VFR transmission, we performed an experiment to compare it with two other methods for reducing the bit rate: (a) reducing the number of poles, and (b) increasing the quantization step size of the LPC coefficients (log-area ratios). Thirty-two stimulus sentences were prepared by passing four utterances (2 sentences × 2 speakers) through eight vocoder systems in a 2×2×2 factorial design; two values were assigned to each of the three parameters: average frame rate, number of poles, and quantization step size. Eight listeners made seven-point category ratings of quality degradation. The results of the experiment show that, of the three methods studied, the VFR technique produced the highest quality at any given transmission rate (or, equivalently, yielded the lowest bit rate for a fixed level of speech quality).

PACS numbers: 43.70.Lw, 43.70.Ep, 43.70.Jt

INTRODUCTION

Even cursory inspection of spectrograms of speech shows that the rate at which the short-term spectrum changes can vary over a wide range. In stressed vowels, or in strident fricatives (s, sh, z, zh), the spectrum may change very little over periods as long as 150–200 msec. During transitions between acoustically different segments, on the other hand, the spectrum may change very rapidly. Variable frame rate (VFR) LPC vocoders take advantage of this variability to reduce their average bit rate. In a typical system, the power spectrum of a 20-msec interval of input speech (a "frame") is modeled every 10 msec, and whenever the spectrum is changing rapidly, every frame that is analyzed is also transmitted. During slowly changing parts of the signal, however, a frame is not transmitted unless it is different from the preceding transmitted frame by more than some threshold. Preliminary tests suggested that VFR transmission could reduce the frame rate to an average of 35 per sec or less, with negligible loss of speech quality (Makhoul *et al.*, 1974). Such a VFR system could operate directly, without any interface, over a time-asynchronous or variable-rate channel, such as the ARPANET. For use over fixed-rate channels, the VFR system must be interfaced to the channel through a tandem of transmit and receive buffers, with associated data-flow control. This introduces additional delay into the transmission path, but recent work (Blackman *et al.*, 1977) has shown that variable-to-fixed rate conversion can be achieved with negligible loss of quality even for delays as short as 80 msec.

The experiment to be described was performed to follow up an unexpected result in an earlier study, in which advantages much smaller than expected were found for VFR transmission. The purpose of the earlier study (Huggins and Nickerson, 1975) was to develop a small set of speech materials, for use in quality rating studies of LPC vocoders. We argued that LPC vocoding can introduce discrepancies between the input and reconstituted speech in several distinct ways, and showed that these produced different effects on perceived quality. LPC vocoding starts by modeling the spectrum of a short waveform interval (e.g., 20 msec) as the response of an all-pole filter. The more coefficients or poles that can be used to define the filter, the more closely the modeled spectrum can approach the speech spectrum. If too few coefficients are used, detail in the speech spectrum is effectively discarded, and cannot thereafter be recovered. Further losses occur as the LPC coefficients are quantized for transmission. Some of the spectral accuracy lost during quantization may be recovered during resynthesis by appropriate smoothing and interpolation algorithms. The foregoing two processes limit the spectral accuracy that can be achieved for a single frame of speech. We have called this "static spectral accuracy."

Each frame of quantized coefficients represents the input-speech spectrum at a particular instant of time. The smaller the intervals between successive analysis frames, the larger the maximum rate of spectral change that can be accurately retained in the reconstituted speech. Thus the frame-analysis interval controls "dynamic spectral accuracy."

The speech materials we developed attempted to target these sources of spectral errors by concentrating within single sentences all phonemes having similar acoustic properties, as shown in Table I. The results of the experiment, which we describe in more detail below, suggest that our attempt was successful. Subjects judging the quality of these sentences, as processed by a variety of vocoders, are in effect able to compare the vocoders with respect to a single source of degradation at a time, which greatly simplifies their task.

In the earlier experiment, the sentences shown in Table I were recorded by 20 speakers, from which a subset of three males and three females were chosen, such that the full range of speaker characteristics found in the group of twenty was retained. The resulting 36 test sentences (6 sentences × 6 speakers) were processed by a

TABLE I. Test sentences.

Phoneme-specific sentences

(1) Why were you away a year, Roy?
(2) *Nanny* may know my meaning.
(3) His vicious father has seizures.
(4) Which tea-party did *Baker* go to?

General sentences

(5) The little blankets lay around on the floor.
(6) The trouble with swimming is that you can drown.

set of twelve simulated vocoders, which used log-area ratios as the transmission parameters. (For the many desirable properties of log-area ratios, see Viswanathan and Makhoul, 1975). After choosing the number of poles (13, 11, or 9), and frame rate, the quantization step size of each system was chosen so as to equate the bit rates of all twelve systems at 2600 bits per sec. Quantization step size varied between 0.2 and 1.75 dB. Seven of the systems used fixed transmission rates of 67, 50, or 40 frames per sec, and the remaining five were VFR systems with average frame rates between 47 and 31 per sec. Pitch and gain were coded in 11 bits, and were transmitted at the frame rate used for the coefficients, in the fixed rate systems, but at a fixed rate of 50/per sec for the VFR systems, to avoid confounding excitation and spectral variables. In the VFR systems, the input speech was analyzed every 10 msec, but the resulting data frame was not transmitted unless the spectral difference between it and the preceding transmitted frame exceeded a threshold. The spectral difference was measured using a log-likelihood ratio measure (Itakura, 1975, Makhoul *et al.*, 1974), and thresholds between 1.0 and 1.75 dB were used. Therefore, frames were sent every 10 msec during rapidly changing parts of the speech, but as seldom as every 80 msec during slowly changing portions. For each of the five VFR systems, the parameter values were chosen so that the average transmission rate over all 36 test sentences was about 2600 bits per sec. The waveform was low-pass filtered at 5 kHz, sampled at 10 kHz, and preemphasized by differencing, before processing through the vocoders.

Subjects rated the degradation of speech quality in each of the 36 stimulus sentences as processed by each of the 12 vocoders. Mean ratings were analyzed by a multidimensional scaling program (MDPREF, see Carroll, 1972), which represents the vocoder systems as points in an N-dimensional space (three dimensional, in our case), and each speaker-sentence combination as a vector through the space. The performance of each vocoder on a particular speaker-sentence combination is represented by the projection of the point representing the system onto the vector representing the stimulus sentence.

The results showed a clear separation of the systems (1) as a function of the number of poles, and (2) as a function of the frame-analysis interval, of the vocoders. Furthermore, the separation along these two dimensions was orthogonal, suggesting that the perceptual effect of changing the number of poles ("static" spectral accuracy) was independent of the perceptual effect of changing the frame-analysis interval ("dynamic" spectral accuracy). The orientation of the test-sentence vectors in the space showed that the separation of the fixed-rate systems by frame-analysis interval was achieved as a result of the specially composed sentence materials (Table I), with the short analysis-interval systems performing better on the rapidly changing sentence [see sentence (4)], and the long analysis-interval systems, with more bits per frame, doing better on the slowly varying sentences [(1) and (2)]. The VFR systems were located correctly for their frame-analysis intervals of 10 msec. Further evidence that our sentences differed in rate of spectral change, as required, is provided by measurements of the average frame rates across the five VFR systems. The average VFR rate was lowest for each of the six speakers in sentences (1) and (2), and highest in sentence (4). Separation of the vocoders as a function of the number of poles resulted from the use of the different talkers, with the relative performance of systems with 13, 11, and 9 poles on a particular sentence being highly correlated with the mean fundamental frequency in the sentence. Nine-pole systems performed almost as well as 11- or 13-pole systems on high-pitched sentences, but much worse on low-pitched sentences.

The five VFR systems included in this study performed less well than expected. Although they did perform better than fixed rate systems on the rapidly changing sentences [sentences (3) and (4) in Table I], they performed worse than some of the fixed-rate systems on the slowly changing sentences, sentences (1) and (2), and about equally well on the general sentences, sentences (5) and (6). On the other hand, the average frame rate of the VFR systems was higher than that of the fixed-rate systems during the rapidly changing sentences, and lower during the slowly changing sentences, which may partly account for the observed performance. At the same time, the large expected advantages of the VFR systems did not appear, and the experiment described below was performed specifically to establish that they do, in fact, occur.

I. PROCEDURE

Equating the bit rates of all vocoders in the earlier study meant that any pair of vocoders differed in at least two parameter values, making comparisons difficult. Therefore for the present study, which had the explicit aim of comparing systems, we adopted a factorial design, in which two values of each of the three parameters occurred in every possible combination. Details of the systems are shown in Table II. These systems represent a much wider range of qualities than was used in the earlier study.

Each system used either 11 or 8 poles. The log-area ratio coefficients were quantized in steps of either 0.5 or 2.0 dB (Viswanathan and Makhoul, 1975). LPC analysis of the speech signal was carried out at 50 frames per sec, and the threshold of the VFR scheme was set to either 0 dB, in which case every analyzed

TABLE II. System parameters for the eight vocoders studied.

System I. D.	PQR	Poles	Quant (dB)	VFR Thresh	Rate F/sec	Bits per sec
A	000	11	0.5	0 dB	50	3157
B	001	11	0.5	2.5 dB	23	1831
C	010	11	2.0	0 dB	50	2155
D	011	11	2.0	2.5 dB	23	1346
E	100	8	0.5	0 dB	50	2521
F	101	8	0.5	2.5 dB	23	1456
G	110	8	2.0	0 dB	50	1771
H	111	8	2.0	2.5 dB	23	1119

frame was transmitted, yielding a fixed frame rate of 50 per sec, or 2.5 dB, which resulted in a variable frame rate that averaged 23.3 per sec. Note that 2.5 dB represents a very coarse threshold, and that the resulting average frame rate is less than 60% of the average frame rate of the VFR systems in the earlier study, over the same sentences. Pitch and gain were coded in 11 bits and transmitted at a constant rate of 50 frames per sec, as in the VFR systems in the earlier study.

A subset of the 36 stimulus sentences used in the first study was selected. To ensure that the subset was representative of the whole set of 36, we chose the two "general" sentences from Table I [i.e., sentences (5) and (6)], since between them these contain (almost) all the phonemes of English. We eliminated the phoneme-specific sentences, since they form a balanced set, and choosing one of them would have entailed choosing the others as well. We then selected two speakers, one male and one female, such that the vectors corresponding to their productions of the two general sentences were separated as widely as possible in the MDPREF solution space of the earlier study. To confirm that these four stimulus sentences were adequately representative, we repeated the MDPREF analysis of the earlier study, using only the subset of data collected on the four sentences. The solution obtained was highly similar to the solution obtained with the whole set of 36 sentences, and achieved the same orthogonal separation of the systems by number of poles, and by frame rate. This test confirmed that the selected subset was indeed representative.

The four sentences were passed through the eight simulated vocoders, and were recorded in two random orders on the stimulus tape, with order of sequential presentation counterbalanced fully across system pairs, and as far as possible across sentence pairs, with the constraint that no system and no sentence should follow itself. Eight subjects were then run individually through two exact repetitions of the tape—although the subjects were not aware of the repetition. Thus each subject made four ratings on each of the 32 stimulus sentences. They rated the degradation of what they heard on a seven-point scale, 1–7, with "overflow bins" (0 and 8) at each end. That is, if a stimulus sounded appreciably better than a previous one labeled with a "1", the subject was allowed to use a "0" response.

II. RESULTS AND DISCUSSION

The mean ratings assigned to the eight systems are shown in Fig. 1, where the ratings are plotted against overall bit rate including pitch and gain. Lines join each pair of systems that differ in only a single parameter: solid lines join all pairs of systems that differ only in frame rate; dashed lines join pairs of systems that differ only in the number of poles; and dotted lines join pairs that differ only in quantization step size.

Consider first the three lines leaving system A, at the upper right hand corner of Fig. 1. For each parameter, system A has the parameter value associated with better speech quality. Bit rate can be reduced for this system in three ways: (1) by reducing the number of poles, (2) by coarsening the quantization, or (3) by going to a VFR system. The figure shows that reducing the number of poles resulted in the smallest savings in bits, accompanied by a large loss of quality. Increasing the quantization step size yielded a slightly better rate of bits saved per unit quality loss. More bits were saved, but at a cost of a slightly larger reduction in quality. Both the largest savings in bits and the smallest drop in quality were associated with the introduction of the VFR scheme. Similar conclusions can be drawn from looking at the gains in quality achieved by increasing the bit rate of the worst system, system H. The smallest quality improvement, with the largest cost in extra bits, was obtained by abandoning the VFR scheme. For one pair of otherwise identical systems, going from fixed to variable frame rate reduced the bit rate by about 40% with no effect on quality (see systems C and D in Fig. 1). All but three of the quality differences, between pairs of systems joined by lines, are extremely significant—that is, well beyond the 0.001 level. The three exceptions were (1) the quality difference between systems C and D, which was not significant; (2) the difference between systems G and H, which just failed to reach significance at the 0.05 level, and (3) the difference between F and H, which was just significant ($P < 0.05$).

Comparison of the variances of the judgments for

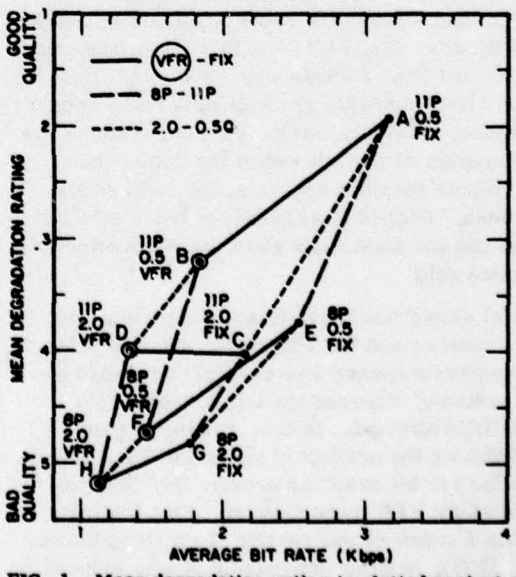


FIG. 1. Mean degradation rating is plotted against average bit rate (including pitch and gain), for each of the eight LPC vocoder systems tested. See text for more details.

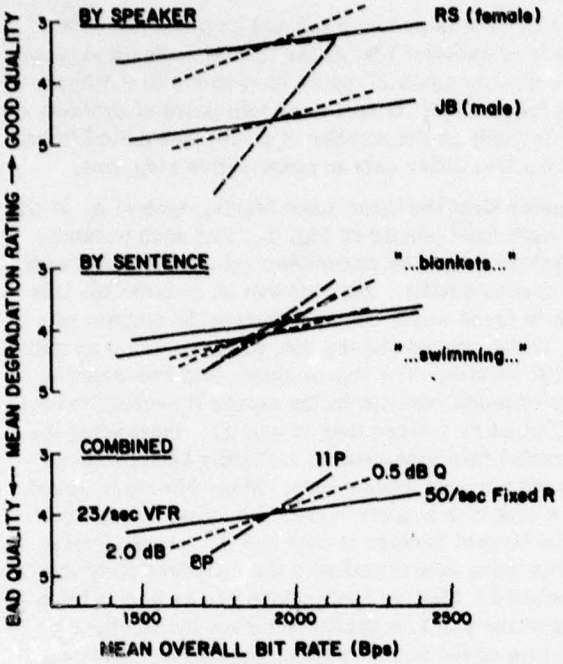


FIG. 2. The effect of speaker, sentence, and vocoder parameter on speech quality. Mean degradation ratings are plotted (a) against mean bit rates for each of the two speakers; (b) for each of the two sentences; and (c) averaged across speakers and sentences. The solid lines connect the points representing the averages for the four systems with fixed frame rate (50 frames per sec) with the points representing the average for the four systems with variable frame rate (23 frames/sec). Similarly, the dashed lines join the means for the four systems using 0.5 dB quantization with those using 2.0 dB. The dotted lines join the means for all the 11-pole systems with those for all the 8-pole systems.

pairs of systems showed that two pairs of systems yielded significant variance ratios. The system pairs are A and E, and B and F, both of which differ only in the number of poles used. The quality judgments for the speech passed through the 8-pole systems (E and F) had a much broader variability—in fact the distribution was bimodal, and Fig. 2 shows why. Here, in Figs. 2(a) and 2(b), the judgments are broken down by speaker and by sentence. Judgments for all 8-pole systems are pooled, and appear at the left end of the dotted lines. Those for 11-pole systems appear at the right end of the dotted lines. Dashed lines show the mean effect of quantization and the solid lines show the mean effect of variable frame rate.

Figure 2(a) shows that there is a strong interaction between the speaker and the effect of number of poles. The male speaker's speech was severely degraded by the 8-pole systems, whereas the female speaker's speech was little affected. In fact, for the female speaker, reducing the number of poles yielded a rate of quality decline per bit saved no greater than that obtained by adopting VFR transmission. This finding corroborates a result of our earlier study (Huggins and Nickerson, 1975), in which we found a strong interaction between the vocoder and the talker's voice in determining speech quality. The relative speech quality of systems using 13, 11, and 9 poles on a particular sentence was highly dependent on the mean fundamental frequency in the test sentence. However, it is likely that the critical variable is not the fundamental frequency, but rather the length of the speaker's vocal tract, which tends to correlate highly with fundamental (large men have low voices). A speaker with a long vocal tract has lower-frequency formants than one with a short vocal tract, so there may be more formants to be modeled within the 5-kHz passband of the vocoder. To separate the effects of fundamental frequency from those of vocal-tract length, one would have to repeat the experiment, using materials that held one constant while varying the other. For example, tract length could be held constant while fundamental was varied, by having single speakers produce each sentence several times, at widely differing pitches. Tract length could be varied, with fundamental held constant by having several speakers, with widely different tract lengths, all produce a sentence at the same fundamental.

III. CONCLUSIONS

Our results confirm that VFR transmission can yield substantial savings in bit rate, with only minor loss of quality. The rate of bits saved, per unit quality loss, is highest for savings achieved by VFR transmission, and lowest for those achieved by reducing the number of poles used in spectral modeling—at least for the parameter values studied here. Secondly, there are major interactions between perceived speech quality and the fundamental frequency of the talker, for some systems.

During the course of this research, it has become clear that the method used to decide whether or not the current data frame should be transmitted is of paramount importance in maintaining good speech quality. The log-likelihood ratio method we used was very simple to implement, and it performed well in rapidly changing sentences, but did not seem to work well in slowly changing sentences. We have recently developed a new VFR scheme (Viswanathan *et al.* 1977), in which log-area ratios are used directly in deciding which frames to transmit, and which explicitly takes into account the linear interpolation performed at the receiver to approximate the coefficients in the frames whose transmission is suppressed. Thus it is sensitive to spectral errors that arise anywhere between two transmitted frames, rather than considering only the end points. This work has demonstrated good quality transmission with average frame rates as low as 26 per sec (and as low as 18 per sec on the slowly changing sentences). Informal listening tests showed that the speech transmitted at 26 frames per sec by the new method was of better quality than that transmitted at 37 frames per sec by the likelihood ratio method.

[a]A condensed version of this paper was presented at the 92nd meeting of the Acoustical Society of America, San Diego, California, Nov 15–19, 1976. The research was supported by the Information Processing Techniques branch of the Advanced Research Projects Agency.

Blackman, E., Viswanathan, R., and Makhoul, J. (1977). "Variable-to-fixed rate conversion of narrowband LPC speech," Proc. Int. Conf. Acoust. Speech Signal Process., Hartford, CT, 9–11 May.

Carroll, J. D. (1972). "Individual differences and multidimensional scaling," in *Multidimensional scaling: Theory and applications in the behavioral sciences*, edited by R. N. Shepard, A. K. Romney, and S. Nerlove (Seminar, New York), Vol. 1, pp. 105–155.

Huggins, A. W. F., and Nickerson, R. S. (1975). "Some Effects of Speech Materials on Vocoder Quality Evaluations," J. Acoust. Soc. Am. 58, s 129(A) (submitted for publication).

Itakura, F. (1975). "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 67–72.

Makhoul, J., Viswanathan, R., Cosell, L., and Russell, W. (1974). *Natural Communication with Computers*, Final Report, Vol. II., Speech Compression Research at BBN, Report No. 2976, Bolt Beranek and Newman Inc., Cambridge, MA, NTIS No. AD/A 003478/5GA, 104 pp.

Viswanathan, R. and Makhoul, J. (1975). "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. Acoust. Speech Signal Process., ASSP-23, 309–321.

Viswanathan, R., Makhoul, J. and Wicke, R. (1977). "The Application of a Functional Perceptual Model of Speech to Variable-Rate LPC Systems," Proc. Int. Conf. Acoust. Speech Signal Process., Hartford, CT 9–11 May.

APPENDIX 11

PHONEME-SPECIFIC

INTELLIGIBILITY TEST

These lists were developed by K. N. Stevens (1962a, b), and have never been published before in their entirety. We thank Professor Stevens for permission to include them here.

K. N. Stevens, M. H. L. Hecker and K. D. Kryter, "An Evaluation of Speech Compression Systems," BBN Report No. 914, March 1962a.

K. N. Stevens, "Simplified Nonsense-Syllable Tests for Analytic Evaluation of Speech Transmission Systems," J. Acoust. Soc. Amer., 34, p. 729, May 1962b.

Lists 1 and 2 (Consonants) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

Lists 3 and 4 (Consonants) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

Lists 5 and 6 (Consonants) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

Lists 7 and 8 (Consonants) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

Lists 9 and 10 (Consonants) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

TEST NO. #AM NAME _____
CONSONANTS: b d m n
VOWELS: I ɛ ʌ ʊ

1. nʌn
2. mɛm
3. mʊm
4. mɪm
5. hɪh
6. nʊn
7. hʌh
8. nɪn
9. dɛd
10. hʊh
11. mɛm
12. hɪh
13. dʊd
14. mʌm
15. nɛn
16. hɛh
17. dʊd
18. dʌd
19. dɪd

TEST NO. #BM NAME _____
CONSONANTS: m n p t
VOWELS: I ɛ ʌ ʊ

1. tʊt
2. mʌm
3. mɪm
4. nʊn
5. pʊp
6. tɪt
7. nɛn
8. pɛp
9. mʌm
10. tʊt
11. nʌn
12. pʌp
13. mɛm
14. nɪn
15. tɛt
16. mʊm
17. tʌt
18. pʌp
19. pɪp

TEST NO. #CM NAME _____
CONSONANTS: f s v z
VOWELS: I ɛ ʌ ʊ

1. zʊz
2. vʌv
3. vɪv
4. fʌf
5. sɛs
6. zɛz
7. vʊv
8. fʊf
9. sʊs
10. sɛs
11. vɛv
12. sɪs
13. sʌs
14. sʌs
15. zɪz
16. fɪf
17. vɪv
18. zʌz
19. fɛf

TEST NO. #AF NAME _____
CONSONANTS: b d m n
VOWELS: I ɛ ʌ ʊ

1. nʊn
2. hʌh
3. dɪd
4. nɛn
5. mʊm
6. nʌn
7. nɪn
8. hʌh
9. dʊd
10. mɛm
11. mɪm
12. mʌm
13. hɛh
14. dʌd
15. dɛd
16. dɛd
17. hʊh
18. nʊn
19. hɪh

TEST NO. #BF NAME _____
CONSONANTS: m n p t
VOWELS: I ɛ ʌ ʊ

1. pɪp
2. mʊm
3. mʊm
4. tʌt
5. pʊp
6. pʌp
7. mɛm
8. nɪn
9. pʌp
10. tʊt
11. nʌn
12. mɪm
13. pɛp
14. tɛt
15. mʌm
16. tɛt
17. nɛn
18. nʊn
19. tɪt

TEST NO. #CF NAME _____
CONSONANTS: f s v z
VOWELS: I ɛ ʌ ʊ

1. vɪv
2. vʊv
3. fɪf
4. zɪz
5. sʌs
6. fʌf
7. zʌz
8. fɛf
9. vɛv
10. sɛs
11. sɛs
12. zʊz
13. fʊf
14. vʌv
15. sʊs
16. zɛz
17. vʊv
18. fʌf
19. sɪs

List 11 (Vowels) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

TEST NO. 12AM NAME _____

CONSONANTS: b d m n
VOWELS: i ɪ ɛ æ

1. d i d
2. m ɛ m
3. m æ m
4. n æ n
5. h ɪ h
6. h i h
7. n i n
8. m ɛ m
9. n ɛ n
10. h ɛ h
11. h æ h
12. d i d
13. d æ d
14. n ɪ n
15. m i m
16. h ɪ h
17. m ɪ m
18. d ɪ d
19. d ɛ d

TEST NO. 12BM NAME _____

CONSONANTS: m n p t
VOWELS: i ɪ ɛ æ

1. n ɪ n
2. m i m
3. p æ p
4. t æ t
5. m ɪ m
6. t æ t
7. n æ n
8. n ɛ n
9. t i t
10. p i p
11. m ɛ m
12. m ɛ m
13. t ɪ t
14. p ɛ p
15. p ɪ p
16. m æ m
17. n i n
18. p ɪ p
19. t ɛ t

TEST NO. 12CM NAME _____

CONSONANTS: f s v z
VOWELS: i ɪ ɛ æ

1. s ɪ s
2. s ɛ s
3. f æ f
4. v i v
5. z ɪ z
6. s ɛ s
7. v æ v
8. f ɛ f
9. s i s
10. f æ f
11. z i z
12. v ɪ v
13. s æ s
14. v ɛ v
15. z æ z
16. f ɪ f
17. z ɛ z
18. z æ z
19. f i f

TEST NO. 12AF NAME _____

CONSONANTS: b d m n
VOWELS: i ɪ ɛ æ

1. h æ h
2. d i d
3. d æ d
4. n ɪ n
5. m i m
6. h ɪ h
7. m ɪ m
8. d ɪ d
9. d ɛ d
10. d i d
11. m ɛ m
12. m æ m
13. n æ n
14. h ɪ h
15. h i h
16. n i n
17. m ɛ m
18. n ɛ n
19. h ɛ h

TEST NO. 12BF NAME _____

CONSONANTS: m n p t
VOWELS: i ɪ ɛ æ

1. t i t
2. n ɛ n
3. n æ n
4. t æ t
5. m ɪ m
6. t æ t
7. p æ p
8. m i m
9. n ɪ n
10. t ɛ t
11. p ɪ p
12. n i n
13. m æ m
14. p ɪ p
15. p ɛ p
16. t ɪ t
17. m ɛ m
18. m ɛ m
19. p i p

TEST NO. 12CF NAME _____

CONSONANTS: f s v z
VOWELS: i ɪ ɛ æ

1. s ɛ s
2. z i z
3. v æ v
4. z æ z
5. s ɪ s
6. f ɪ f
7. z ɪ z
8. f æ f
9. s i s
10. z ɛ z
11. f i f
12. v æ v
13. f i f
14. v i v
15. f ɛ f
16. s ɛ s
17. s æ s
18. v ɪ v
19. v ɛ v

List 12 (Vowels) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

13AM
TEST NO. _____ NAME _____
CONSONANTS: b d m n
VOWELS: u ʊ ʌ ɑ

1. b u b
2. n ʊ n
3. b ʊ b
4. d ɑ d
5. d ɑ d
6. d ʌ d
7. b ɑ b
8. m ʌ m
9. m u m
10. m ɑ m
11. d ʊ d
12. b ʌ b
13. n u n
14. n ʌ n
15. m ʊ m
16. n ɑ n
17. d u d
18. b ʌ b
19. n ʊ n

13BM
TEST NO. _____ NAME _____
CONSONANTS: m n p t
VOWELS: u ʊ ʌ ɑ

1. p ʊ p
2. n u n
3. m ɑ m
4. p ʊ p
5. p ʌ p
6. t ʊ t
7. m ʌ m
8. m ʌ m
9. p u p
10. t u t
11. n ʌ n
12. n ɑ n
13. t ɑ t
14. m ʊ m
15. t ɑ t
16. p ɑ p
17. m u m
18. n ʊ n
19. t ʌ t

13CM
TEST NO. _____ NAME _____
CONSONANTS: f s v z
VOWELS: u ʊ ʌ ɑ

1. f u f
2. v ɑ v
3. f u f
4. v u v
5. f ʌ f
6. s ʌ s
7. z ɑ z
8. v ʊ v
9. v ʌ v
10. s ʌ s
11. z u z
12. v ɑ v
13. z ɑ z
14. s ʊ s
15. f ʊ f
16. z ʊ z
17. f ɑ f
18. s u s
19. z ʌ z

13AF
TEST NO. _____ NAME _____
CONSONANTS: b d m n
VOWELS: u ʊ ʌ ɑ

1. d ɑ d
2. d ɑ d
3. b ʊ b
4. n ʊ n
5. b u b
6. n ʊ n
7. b ʌ b
8. d u d
9. n ɑ n
10. m ʊ m
11. n ʌ n
12. n u n
13. b ʌ b
14. d ʊ d
15. m ɑ m
16. m u m
17. m ʌ m
18. b ɑ b
19. d ʌ d

13BF
TEST NO. _____ NAME _____
CONSONANTS: m n p t
VOWELS: u ʊ ʌ ɑ

1. p ʌ p
2. t ʊ t
3. m ʌ m
4. m ʌ m
5. p u p
6. t u t
7. n ʌ n
8. n ɑ n
9. t ɑ t
10. m ʊ m
11. t ɑ t
12. p ɑ p
13. m u m
14. n ʊ n
15. t ʌ t
16. p ʊ p
17. n u n
18. m ɑ m
19. p ʊ p

13CF
TEST NO. _____ NAME _____
CONSONANTS: f s v z
VOWELS: u ʊ ʌ ɑ

1. z ʊ z
2. f ɑ f
3. s u s
4. z ʌ z
5. f u f
6. v ɑ v
7. f u f
8. v u v
9. f ʌ f
10. s ʌ s
11. s ɑ s
12. v ʊ v
13. v ʌ v
14. s ʌ s
15. z u z
16. v ɑ v
17. z ɑ z
18. s ʊ s
19. f ʊ f

List 13 (Vowels) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

NAM
TEST NO. ____ NAME _____
CONSONANTS: b d m n
VOWELS: i æ a u

1. mæm
2. bab
3. bæd
4. dad
5. did
6. bib
7. nin
8. bub
9. nid
10. bæb
11. dud
12. nan
13. mim
14. næn
15. nun
16. bæb
17. did
18. mam
19. mum

NBM
TEST NO. ____ NAME _____
CONSONANTS: m n p t
VOWELS: i æ a u

1. pap
2. nin
3. mæm
4. pap
5. pup
6. tat
7. mum
8. mum
9. pip
10. tit
11. nun
12. næn
13. tæt
14. mam
15. tæt
16. pæp
17. mim
18. nan
19. tut

NCM
TEST NO. ____ NAME _____
CONSONANTS: f s v z
VOWELS: i æ a u

1. vuv
2. faf
3. sis
4. viv
5. vuv
6. fif
7. zaz
8. sas
9. saf
10. zaz
11. fæf
12. væv
13. sæs
14. sæs
15. zuz
16. fuf
17. vav
18. sus
19. zæz

MAF
TEST NO. ____ NAME _____
CONSONANTS: b d m n
VOWELS: i æ a u

1. mum
2. mam
3. did
4. bæb
5. nun
6. næn
7. mim
8. nan
9. dud
10. bæb
11. nin
12. bub
13. nin
14. bib
15. did
16. dad
17. dæd
18. bab
19. mæm

NBF
TEST NO. ____ NAME _____
CONSONANTS: m n p t
VOWELS: i æ a u

1. mum
2. pip
3. tit
4. nun
5. næn
6. tæt
7. mam
8. dæd
9. pæp
10. mim
11. nan
12. tut
13. pap
14. nin
15. mæm
16. pap
17. pup
18. tat
19. mum

NCF
TEST NO. ____ NAME _____
CONSONANTS: f s v z
VOWELS: i æ a u

1. zæz
2. sus
3. vav
4. fuf
5. zuz
6. sæs
7. sæs
8. væv
9. fæf
10. zaz
11. faf
12. sas
13. zaz
14. fif
15. vuv
16. viv
17. sis
18. faf
19. vuv

List 14 (Vowels) from the Phoneme-Specific Intelligibility Test (Stevens, 1962a, b)

APPENDIX 12

A FRAMEWORK FOR THE OBJECTIVE EVALUATION OF VOCODER SPEECH QUALITY

(Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, April 1976.)

A FRAMEWORK FOR THE OBJECTIVE EVALUATION OF VOCODER SPEECH QUALITY

John Makhoul
R. Viswanathan
William Russell

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

While there exist methods in the literature for objectively evaluating the intelligibility of speech in the presence of stationary noise, little has been done regarding the objective evaluation of either the intelligibility or the quality of vocoded speech. We present a framework within which we have begun a step-by-step program to develop objective measures for vocoded speech quality that are consistent with results from subjective tests.

1. Introduction

The ultimate criterion for determining the quality of the speech that is produced by any compression, encoding or transmission system is the way it sounds to the human listener. Although there are well established procedures to test the intelligibility of speech, little work has been done in developing procedures to test speech quality, and in particular vocoder speech quality. The few procedures that are available are subjective and require extensive testing with human listeners, which is expensive in terms of both time and money.

It would be desirable to develop objective procedures for speech quality evaluation that correlate well with the scores obtained from subjective listening tests. These objective measures would ensure uniformity in evaluation as well as enable the evaluation to be done by computer. Also, the measures can be used in the design of better quality vocoders. While there exist methods in the literature for objectively evaluating the intelligibility of speech in the presence of stationary noise [1,2], little has been done regarding the objective evaluation of either the intelligibility or the quality of vocoded speech. The problem is that if one regards the distortion in the vocoded speech signal as noise superimposed on the signal, then this noise is not only nonstationary but is correlated with the signal. This makes the problem of objective evaluation of vocoded speech quality a difficult one. However, given the immense long-term benefits in terms of time and expense, any headway into the solution of the problem is desirable.

This paper presents a framework within which we have begun a step-by-step program to develop objective measures of vocoded speech quality that are consistent with results from subjective tests.

2. Necessary Conditions

Let s(n) be the original speech signal and s′(n) a vocoded version of the same signal. Our aim is to develop measures that compare the quality of s′(n) relative to s(n). Note that the formulation of this problem is different from that of the objective evaluation of speech intelligibility in the presence of noise. In the latter, the noise spectrum is assumed stationary and can be measured directly. The resulting objective intelligibility scores are obtained by comparing the average signal spectrum to the noise spectrum [1,2]. The same procedure cannot be applied in the case of vocoded speech because the "noise" that corrupts the signal is not well defined, and in any case not easily measured. Even if the latter were possible, such noise cannot be considered stationary and it is also correlated with the signal. Therefore, one must somehow compare the vocoded signal s′(n) to the original signal s(n).

One of the main problems in comparing s′(n) to s(n) is that of time synchronization, so that corresponding segments of the two signals can be compared. However, assuming that somehow one is able to align the two signals, the problem of comparing s′(n) to s(n) remains.

In many communication systems, the average mean squared difference error between two signals is taken as a measure of distance or deviation between the two signals. It is simple to show that such an error measure cannot be a measure of the difference in quality between the two signals. This is done by offering a counterexample. Let s(n) be the input to an all-pass filter, and let s′(n) be its output. The filter can be designed such that the wave shape of s′(n) is quite different from s(n), and such that the mean squared difference between s′(n) and s(n) is large. However, we know from perceptual experiments that, in all likelihood, the difference between s′(n) and s(n) is insignificant as judged by a human listener (at least for vocoder purposes). In fact, it is well known that, except for pitch, phase information is quite irrelevant to the perception of speech [3]. It is difficult to imagine an error criterion on the waveform which would be insensitive to phase.

The answer is clearly to go to the spectrum. In fact, vocoders have traditionally transmitted parameters related to the magnitude of the spectrum. Channel vocoders have used one type of phase realization for synthesis, and LPC vocoders have used another (minimum phase). The problem, then, seems to reduce to a comparison between the short-time spectra of s′(n) and s(n). But the spectrum is only one aspect of the signal that is important to perception and is distorted by the vocoder. The other important aspect is the source information.

After some thought it became clear to us that objective measures for the evaluation of vocoded speech quality must obey two maxims: (1) They must be a function of the vocoding

process, and in particular the vocoder transmission parameters,

(2) They must somehow relate to perception.

The first maxim basically says that the objective evaluation of vocoded speech quality cannot be done abstractly, treating $s'(n)$ as some arbitrary distortion of $s(n)$, but rather it must relate directly to the vocoding process. The second maxim merely states the obvious necessity to have the objective measures be perceptually meaningful. These two maxims not only form a sound basis on which to build these measures, but also offer the hope of a diagnostic tool for the evaluation and refinement of vocoder design. Based on the two maxims, therefore, we proceeded to develop the general framework for objective quality evaluation.

3. Determiners of Quality

In a vocoder system, there are four major identifiable components that can contribute to the degradation of vocoded speech quality: analysis, encoding, transmission, and synthesis. We shall discuss the types of errors introduced by the different components, in an effort to identify the major determiners of speech quality in the vocoding process. This would then give us a handle with which to design our objective measures.

Transmission

Channel transmission errors are an important factor in the choice of a vocoder system, in that different vocoders are affected differently by different types of channel errors. However, given that error correcting codes can reduce sharply the effective error rate, one must still explain the degradation in quality due to the vocoder itself. Therefore, in attempting to develop objective quality measures, we shall assume that channel transmission errors are negligible.

Analysis

The importance of the analysis component is apparent when we consider that it embodies the particular speech model employed. The parameters extracted in this component determine the upper bound on the quality of the synthesized speech.

The general vocoder speech model is that of a source exciting a system that represents the short-time spectrum. We shall restrict our discussion here to LPC vocoders, with the knowledge that it can be extended easily to other types of vocoders (e.g. channel vocoders). The LPC model is that of a source with a relatively flat spectral envelope, exciting an all-pole filter. There are three main types of LPC vocoders, depending on the type of source excitation: residual excited, voice excited, and pitch excited. However, all three types of vocoders perform essentially the same type of analysis to obtain the filter parameters. Although there may be speech quality differences depending, for example, on whether the covariance, autocorrelation or lattice method of linear prediction is used, these differences tend not be of a major nature. The upper bound on the vocoded speech quality is basically a function of the type of excitation used. This is discussed below for each of the three types of LPC vocoders.

Residual Excited Vocoder. In this type of vocoder [4], the residual signal is used to excite a filter that is the exact inverse of the filter used to generate the residual signal from the speech signal. Assuming no quantization errors in either the residual signal or the filter parameters, the synthesized signal $s'(n)$ will be almost identical to the original signal $s(n)$. Therefore, here, the analysis itself does not degrade the speech quality.

Voice Excited Vocoder. In this type of vocoder [5,6], a down sampled baseband comprises the source information. At the receiver the baseband is nonlinearly processed to obtain an excitation function with a flat spectrum. Even under no parameter quantization, the synthesized signal $s'(n)$ will be different from $s(n)$. Therefore, the speech model employed is already responsible for a certain change in the speech quality when compared to the original. One method of estimating this change in quality would be to compare the filter excitation signal for this vocoder to the residual signal used in the residual excited vocoder. Such comparison is probably not straightforward, but it is made easier by the fact that the two signals are more or less time-synchronized (in terms of where pitch periods are, etc.).

Pitch Excited Vocoder. In this case, the excitation is either a sequence of pitch pulses or white noise. Here, $s'(n)$ resembles $s(n)$ in its gross features, but certainly not in the detailed signal structure. Also, unlike the voice excited case, $s'(n)$ is generally not synchronized with $s(n)$, because the voiced/unvoiced (V/UV) excitation is not synchronized with the residual signal, which makes it difficult to get an objective estimate of the change in quality due to the pitch excited model. This is unfortunate considering that the V/UV decision is perhaps the single most important one that affects the quality of $s'(n)$. There are currently no established procedures for the automatic evaluation of V/UV decisions. The existing procedures are manual, in that intervention by a human is necessary to establish whether a voiced or an unvoiced decision would be appropriate for each frame in the analysis (and whether the extracted pitch value is accurate). In certain critical situations, such decisions are made by trial and error as to which sounds better. There are other cases where a mixed voiced-frication source is more appropriate. Thus far, these cases have not been dealt with successfully in vocoders.

Because of the dearth of good testing procedures to evaluate the effects of the excitation on speech quality, we have decided to table this problem in our initial search for objective measures of quality.

Synthesis

Although a large part of the synthesis process is dictated by the type of model used and signal analysis performed, there remain a number of design choices in the synthesizer

that can noticeably affect the synthesized speech quality. The major choices relate to the updating and interpolation of filter parameters, as well as the choice of the filter implementation structure. For example, we have found that if the analysis is performed time-synchronously, it is best to interpolate and update filter parameters time-synchronously as well [7].

Although there are important issues relating to filter implementation structure (for example, placing the gain at the output of a normalized filter [8] causes "clicks" to occur during large changes in gain), it is always possible to choose the implementation structure in such a way that the structure itself contributes negligibly to the degradation of the quality.

Encoding

We include in this component
(1) the choice of the number of parameters to transmit,
(2) how to quantize them, and
(3) when to transmit them.
The parameters include the source parameters (the residual signal in a residual excited vocoder, or pitch and gain in a pitch excited vocoder), and the synthesizer parameters, which can take different forms, with the most popular being the log area ratios in an LPC vocoder [9], or the output energies of the channel filters in a channel vocoder. The choice of the number of parameters, along with their quantization, determine to a large extent the static signal quality at specific time instances, while the transmission and update rate determine the dynamic signal fidelity.

Conclusion

For narrow-band vocoder systems (less than 5000 bps), the encoder, as we have defined it, is the major determiner of speech quality. This is due to the heavy quantization that is necessary to produce low bit rates. Design issues in the analysis and synthesis are important, but for low rate systems, the encoder plays the major role.

4. General Framework

It follows from the previous section that, if the bulk of the synthesized speech quality is determined by the encoder, then one should be able to obtain an approximate objective measure of the quality difference between the original and vocoded speech by somehow comparing the parameter values at the input and output of the encoder. One could also include the interpolation in the synthesis component, and compare the parameter values at the synthesizer with the parameters at the input to the encoder (which are produced by the analysis). In any case, the problem is thus reduced from comparing the quality of two speech signals to comparing two sets of parameters that are related to each other in a well specified manner. This, in turn, implies that such comparisons or quality measuring procedures are to be built "inside" the vocoder instead of outside it. Comparisons are made between the unquantized parameters (reference system) and the parameter values used at the synthesizer (test system).

Inherent in the above analysis is that speech synthesized using the input parameters to the encoder is of very good quality. This is not difficult to achieve. For example, in an LPC vocoder, if the signal bandwidth is 5 kHz, then a 14-pole analysis every 10 ms would give unquantized parameters, which when used in the synthesis, would result in speech whose quality is very good compared to the original speech. This does not necessarily mean that the encoder has to quantize the 14 filter parameters and transmit them every 10 ms. The restriction is merely on the analysis. The encoder may then choose a smaller (and perhaps variable) order for transmission, and at a lower (and perhaps variable) rate [7].

We now state the three observations (assumptions) that form the basis for our work in developing objective quality evaluation measures:
(1) Speech synthesized from unquantized parameters, extracted every 10 ms, is of very good quality compared to the original speech.
(2) Except for pitch and gain, the fidelity of the short-time spectrum is the principal determiner of quality.
(3) The spectrum is uniquely defined by the filter parameters.

The first observation gives us an anchor point defined in terms of the system parameters and against which to compare quantized realizations of the same utterance. The second and third observations relate the filter parameters to speech quality through the concept of spectral fidelity. This, then, gives us a framework within which to develop the desired objective measures of speech quality.

5. An Initial Experiment

Given a speech utterance processed by an LPC vocoder, an objective measure summarizes the error or deviation between the reference and the test sets of parameters in terms of a single number which we shall call an objective evaluation score. The objective score would be expected to reflect the perceived quality (relative to the reference) of the speech utterance if, indeed, the objective measures were sensitive to all quality-determining factors. It is unreasonable, and perhaps too simplistic, to expect that one objective measure could always correctly predict perceived speech quality. The chance of such a prediction may be enhanced by combining a number of objective measures in some fashion to obtain an overall objective score. Each measure may be sensitive to some aspects of quality. Ultimately, we plan to perform a multidimensional analysis [11] on the objective scores obtained from a number of measures with the hope of relating them to different quality dimensions yet to be discovered. For the present study, however, we chose to develop a number of objective measures and investigate each of them separately so as to become familiar with their properties.

For each data frame, an error between the reference and the test parameters is computed using an appropriate "distance" measure. Ideally, such frame errors should be computed only at selected points in time within the speech utterance that are "perceptually significant." For the purposes of the present study, we simply computed the frame error at a fixed rate, say, every 10 ms. We thus had two problems. (1) To develop suitable distance measures to compute frame errors. (2) To combine all the frame errors within a speech utterance into one number, which provides the objective score.

We considered several distance measures for computing the error between the reference and test parameters of a given frame. These measures were based on the power spectrum of the all-pole linear prediction filter. Traditional mean squared differences between log spectra, as well as other measures were used. The errors were also frequency weighted in different ways, including a weighting based on the articulation index [10]. The resulting error sequence at each frame was then combined to give the overall objective score. The sequence was time weighted using the filter gain and the "spectral difference" (rate of change of spectrum) between frames.

These objective measures were used in an initial experiment to correlate the objective scores with the results of a rank ordering experiment of subjective quality that compared different vocoder systems [11]. Different combinations of objective measures were used in the experiment. Comparisons of the objective and subjective scores indicated that no single objective measure was able to always predict correctly the subjective rank ordering of vocoded speech utterances. Furthermore, the objective scores were heavily clustered (relative to the subjective scores) for the different systems, indicating a lack of separability.

6. Program for Research

Based on our initial experiments it became clear that what we need is a step-by-step program to understand the different aspects of the relations between spectral variations and speech quality, in order to be able to begin developing the desired objective measures of quality. First, we shall attempt to discover the quality determining factors in the spectrum independent of time. Following that, we shall attack the more difficult problem of discovering the time-dependent quality determining factors.

As a first step, we have begun to develop spectral distance measures that are consistent with published perceptual data on vowel difference limens. This work is described in a separate paper [10]. One of the important conclusions there is that traditional distance measures between log spectra are not consistent with perceptual data.

7. Conclusions

In this paper we presented the rationale behind the general framework for the objective evaluation of vocoder speech quality. The framework calls for inserting these objective measures inside the vocoder to compare the sets of filter parameters after analysis and before synthesis, in order to observe the effects of encoding and interpolation on the resulting spectra. Spectral variations, in turn, are related to speech quality. A step-by-step program has been initiated to discover the time-independent as well as time-dependent quality determining factors in the short-time spectrum.

Acknowledgment

This work was supported by the Information Processing Techniques Branch of the Advanced Research Projects Agency.

References

1. L.L. Beranek, Acoustics, New York: McGraw-Hill, 1954.

2. K.D. Kryter, The Effects of Noise on Man, New York: Academic Press, 1970.

3. J.L. Flanagan, Speech Analysis Synthesis and Perception, Second Edition, New York: Springer-Verlag, 1972.

4. C.K. Un and D.T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate below 9.6 kbits/s," IEEE Trans. Comm., Vol. COM-23, 1466-1474, Dec. 1975.

5. C.J. Weinstein, "A Linear Prediction Vocoder with Voice Excitation," EASCON '75, Washington, D.C., 30A-30G, Sept. 29-Oct. 1, 1975.

6. B. Atal, M. Schroeder and V. Stover, "Voice-Excited Predictive Coding System for Low Bit-Rate Transmission of Speech," Int. Conf. Comm., San Francisco, Ca., June 1975.

7. J. Makhoul, R. Viswanathan, L. Cosell and W. Russell, Natural Communication with Computers, Final Report, Vol. II, Speech Compression Research at BBN, Report No. 2976, Bolt Beranek and Newman Inc., Cambridge, Mass., Dec. 1974.

8. A. Gray, Jr., and J. Markel, "A Normalized Digital Filter Structure," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 268-277, June 1975.

9. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 309-321, June 1975.

10. R. Viswanathan, J. Makhoul and W. Russell, "Towards Perceptually Consistent Measures of Spectral Distance," IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, April 1976.

11. A.W.F. Huggins and R.S. Nickerson, "Some Effects of Speech Materials on Vocoder Quality Evaluations," J. Acoust. Soc. Am., Vol. 58, Supplement No. 1, S129, Fall 1975.

APPENDIX 13

TOWARDS PERCEPTUALLY CONSISTENT MEASURES OF SPECTRAL DISTANCE

(Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, April 1976.)

TOWARDS PERCEPTUALLY CONSISTENT MEASURES OF SPECTRAL DISTANCE

R. Viswanathan
John Makhoul
William Russell

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

This paper considers distance measures for determining the deviation between two smoothed short-time speech spectra. Since such distance measures are employed in speech processing applications that either involve or relate to human perceptual judgment, the effectiveness of these measures will be enhanced if they provide results consistent with human speech perception. As a first step, we suggest Flanagan's results on difference limens for formant frequencies as one basis for checking the perceptual consistency of a measure. A general necessary condition for perceptual consistency is derived for a class of spectral distance measures. A class of perceptually consistent measures obtained through experimental investigations is then described, and results obtained using one such measure under Flanagan's test conditions are presented.

1. Introduction

Given two smoothed short-time speech spectra, a fundamental problem in speech processing is to determine the distance or the amount of deviation between the two spectra. In speech recognition, the two spectra may correspond to two different speech sounds, or perhaps two different versions of the same sound [1-3]. In speaker verification or identification, the two spectra may correspond to speech produced by either two different speakers or by the same speaker on two different occasions [4,5]. In variable frame rate speech compression, two adjacent analysis frames may have produced the two spectra [6,7]. In the problem of objective evaluation of vocoded speech quality, which the authors have recently formulated [8], the two spectra may correspond to the quantized and the unquantized sets of filter parameters. Still another application of spectral distance measures is in the spectral sensitivity analysis needed for optimal parameter quantization [9].

These examples clearly bring out the importance of spectral distance measures in speech processing. The extent to which a distance measure is valid greatly determines the efficiency of the underlying task in which it is employed. Inasmuch as one strives to achieve a machine performance that is close to what a human can do under the same situation (e.g., first two applications above), or inasmuch as the vocoded speech is to be perceived by human listeners, it is appropriate to require of these distance measures to be at least consistent with the known results of human perception. The work reported in this paper represents a first step towards obtaining perceptually consistent measures of spectral distance.

About two decades ago Flanagan reported perceptual results for determining difference limens for formant frequencies of vowels [10]. One of his results is particularly relevant to this paper. Briefly, when two formants are in close proximity, human perception exhibits an asymmetrical pattern in that moving one of the two formants closer to the other by a given amount produces a larger perceived quality difference than moving that formant away from the other by the same amount. On the other hand, the same formant shifts produce a symmetrical pattern when the two formants are well separated. We use this result as one basis for checking the perceptual consistency of spectral distance measures.

Smoothed spectra can be obtained by using a number of methods such as filter bank, cepstrum, and linear prediction (LP). For simplicity, we focus in this paper on LP spectra, although most of the discussions presented below apply to other types of spectra as well. The LP spectrum is given by [11]

$$P(\omega) = \frac{G^2}{S(\omega)} = \frac{R_o V_p}{\left|1+\sum_{k=1}^{p} a_k e^{-j\omega k}\right|^2} , \quad (1)$$

where G is the linear predictor gain, R_o is the speech signal energy, V_p is the normalized prediction error, $S(\omega)$ is the spectrum of the inverse filter and a_k, $1 \le k \le p$, are the predictor coefficients.

2. Spectral Distance Measures

Let $d(X,Y)$ denote the distance or deviation between the spectra $X(\omega)$ and $Y(\omega)$. From a mathematical viewpoint, one may be tempted to insist that the distance measure satisfy the three axioms of a metric:

(a) Positive definiteness: $d(X,Y) \ge 0$, $d(X,Y)=0$ iff $X \equiv Y$;
(b) Symmetry: $d(X,Y)=d(Y,X)$;
(c) Triangle inequality: $d(X,Y) \le d(X,Z)+d(Z,Y)$.

We require, however, only the property (a) to be true. There are many examples in real life where distance symmetry does not hold. There is no evidence to support the validity of a symmetrical distance in the context of human speech perception. For a similar reason, we do not insist that the property (c) be necessarily true. We postulate that if a distance measure is perceptually consistent, it will prove to perform better in applications involving, or relating to, human perception.

Normalization

Before we define a measure of distance between two LP spectra $P_1(\omega)$ and $P_2(\omega)$, it may be desirable to normalize these spectra in some fashion. For instance, they may be normalized to have the same arithmetic mean (AM) or total energy. Alternately, they may

be normalized to have the same geometric mean (GM), i.e., the log spectra will have the same average.

Error Definition

An error function between the normalized spectra can be defined either in the (linear) spectral domain as

$$e(\omega) = P_1(\omega) - P_2(\omega) \quad , \tag{2}$$

or, in the log spectral domain as

$$e(\omega) = \log P_1(\omega) - \log P_2(\omega) \quad . \tag{3}$$

Other reasonable error definitions include

$$e(\omega) = \left[P_1(\omega) - P_2(\omega)\right]/P_1(\omega) \quad , \tag{4}$$

$$e(\omega) = P_1(\omega)/P_2(\omega) \quad . \tag{5}$$

Spectral Distance Measure

A large class of spectral distance measures can be defined as the weighted L_k norm:

$$d_k(P_1,P_2,W) = \left[\frac{\int_{-\pi}^{\pi} W(P_1(\omega),P_2(\omega),\omega)\,|e(\omega)|^k d\omega}{\int_{-\pi}^{\pi} W(P_1(\omega),P_2(\omega),\omega)\,d\omega}\right]^{\frac{1}{k}} \tag{6}$$

where the weighting function W in general depends on $P_1(\omega)$, $P_2(\omega)$ and frequency ω, and takes only positive values. If the error is defined as in (4) or (5), the distance measure in (6) is not symmetric. Also, if the weighting function depends explicitly on P_1 and P_2, the resulting distance measure is in general not symmetric. In all other cases, a symmetric distance measure results. In the absence of any weighting, d_{-1} is the harmonic mean, d_0 is the GM, d_1 is the AM, and d_2 is the root mean square value of the absolute error function. Between the minimum $d_{-\infty}=\mathrm{Min}|e(\omega)|$ and the maximum $d_{\infty}=\mathrm{Max}|e(\omega)|$, d_k is a monotonically increasing function of k.

The weighting function W in (6) is used to differentially weight individual errors and is determined based on some concept of speech perception. Notice that any constant multiplicative factor in the weighting function does not affect the distance measure. Some specific weighting functions are discussed in Section 4.

Examples: References [2,6,7] use d_2 with the error defined as in (5). (With a Gaussian assumption, this measure becomes a likelihood ratio [2].) Reference [9] employs d_2 with the error given by (3). Cepstral distance measures used in [1,4] have been shown to be highly correlated to d_2 with the error as in (3) [12].

3. A Necessary Condition for Perceptual Consistency

Fig. 1 shows two plots of spectral deviation or distance versus frequency shift of the second formant causing that spectral deviation. (Frequencies of the other three fixed formants and the nominal value of the second formant frequency are given in the figure. Fixed bandwidths of all the formants are as in [10].) Fig. 1(a) corresponds to the error definition (5) while Fig. 1(b) corresponds to the error definition (3). Both plots were obtained using GM normalization, k=1 and no weighting in (6). (We have plotted log d for plot (b) so that ordinates of both plots are in decibels.) The almost symmetrical plots in Fig. 1 do not conform with properties given by Flanagan (see Fig. 4(c) in [10]).

Notice that the two distance measures that produced the plots in Fig. 1 depend only on the ratio of the spectra P_1 and P_2 (in view of (3) and (5)). Below we prove that with GM normalization, any distance measure which is a function of only the ratio of the spectra is necessarily perceptually inconsistent. First, we give our working definition of perceptual consistency, based on Flanagan's results [10].

Working Definition of Perceptual Consistency: Let X and Y be two vowel spectra, such that Y is identical to X except that one of the formant frequencies F is shifted by a variable amount ΔF. A given spectral distance measure d(X,Y) between X and Y is said to be perceptually consistent if

(a) when F is close to another formant F′, d(X,Y) exhibits asymmetry such that it is greater when F is moved ΔF towards F′, than when F is moved ΔF away from F′;
(b) such asymmetry decreases as F and F′ are further apart.

Now, consider a class D_g of spectral distance measures defined by (6) where the error e(ω) is computed after GM normalization of the spectra. For this class of distance measures, a necessary condition for perceptual consistency is provided below in the form of a theorem.

Theorem: A necessary condition for any spectral distance measure $d(P_1,P_2)$ in the class D_g to be perceptually consistent (as defined above) is that it not be a function of only the ratio of the two spectra P_1 and P_2.

Proof: Assume that a distance measure in D_g violates the necessary condition. We show that this distance measure is not perceptually consistent. Let P_2 be obtained from P_1 by shifting only one of its formant frequencies while keeping all other parameters intact. Let the denominator $S(\omega)$ in (1) be factored into $R(\omega)$ and $S'(\omega)$, where $R(\omega)$ is the contribution to the spectrum from the formant under consideration and $S'(\omega)$ represents the contributions from all other poles of the linear predictor. Thus, $P_1(\omega) = 1/(R_1(\omega).S'_1(\omega))$ and $P_2(\omega) = 1/(R_2(\omega).S'_1(\omega))$, where $R_2(\omega)$ is the perturbed version of $R_1(\omega)$. This gives the result that the ratio of P_1 and P_2 depends only on the formant under consideration. Specifically, the ratio does not depend on whether or not this formant is in close proximity to another formant. This clearly establishes that the measure is not perceptually consistent according to our working definition.

With other types of spectral normalization, the ratio of gain terms (G^2 in (1)) of the two spectra depends in general on the overall shape of the spectrum. For instance, with AM normalization, this ratio is between the normalized prediction errors (V_p in (1)) corresponding to the two spectra, which depend on the total spectral shapes [3]. Establishing a general necessary condition for perceptual consistency in these cases is difficult. However, with AM normalization, our experimental results show that when the necessary condition stated above is violated, perceptual consistency is not obtained.

We do not wish to state that perceptually inconsistent measures are not useful. In fact, in the applications mentioned in the introduction, many such measures have been succesfully used. We suggest, however, that use of perceptually consistent measures in these applications may lead to an improved performance of the underlying tasks.

4. Weighting Functions

We have investigated a number of reasonable frequency weighting functions [13]. A brief discussion of some of these weighting functions is given below.

Spectral Intensity Weighting

Since formant peaks of a spectrum are perceptually important, it is reasonable to emphasize spectral errors that occur close to formant peaks. One way of achieving this error weighting is to use $P_1(\omega)$, $P_2(\omega)$, or some generalized mean of the two as weighting functions.

Frequency Derivative Weighting

An alternate method of emphasizing spectral errors that occur close to formant peaks is to employ a suitable function of first and second derivatives of $P_1(\omega)$ or $P_2(\omega)$ for weighting the errors.

Articulation-Index (AI) Based Weighting

AI is a physical measure that is highly correlated with subjective speech intelligibility results. Since it is not unrealistic to consider speech intelligibility and quality as related phenomena, we have derived, by adapting some of the results used in AI computation, a weighting function which decreases exponentially with frequency: $W=\exp(-\alpha\omega)$, where α is a particular constant [13].

All the spectral distance measures that we investigated, even with the use of the above weighting functions, had one common problem in that for the case when the first formant frequency was shifted about the nominal value of 300 Hz, a given amount of left shift always produced a larger spectral deviation than a right shift of the same amount, which is just the opposite of what Flanagan reported (see Fig. 3(a) in [10]). (We found, however, that some of these measures and weighting functions produced the right types of asymmetry in other test conditions considered by Flanagan.) To attempt to overcome this problem, we investigated the following weighting functions based on perceived loudness.

Perceived Loudness Weighting

Based on the work of Stevens [14], we define the perceived loudness function $L(\omega)$ of a spectrum $P(\omega)$ as $[P(\omega)A(\omega)]^{1/3}$, where $A(\omega)$ is shown plotted in Fig. 2. Notice the sharp change of $A(\omega)$ at low frequencies, which may be used to our advantage to overcome the problem mentioned above. The weighting function may then be defined in terms of $A(\omega)$ or $L(\omega)$. We have investigated the following weighting functions: $W=A(\omega)$; $W=L_1(\omega)$ (perceived loudness of $P_1(\omega)$); $W=L_2(\omega)$; $W=|L_1(\omega)-L_2(\omega)|$. Only the weighting function $W=A(\omega)$ produced the right asymmetry for the case when the first formant frequency was shifted about its nominal value of 300 Hz.

In the next section, we give examples of perceptually consistent distance measures which use the weighting function $A(\omega)$.

5. A Class of Perceptually Consistent Distance Measures

Our experimental investigations have led to a class of spectral distance measures which produce the right types of asymmetry attributable to formant interaction under all test conditions considered by Flanagan. This class is defined by (6) with GM normalization, the spectral error defined in the (linear) spectral domain as in (2), and the weighting function $A(\omega)$ shown in Fig. 2.

Figs. 3-5 show plots of spectral distance versus formant frequency shift under three different test conditions for the above measure with $k=1$ in (6). These plots compare rather nicely to the corresponding ones that Flanagan has given. Notice that while our spectral distance plots in general have a monotonically increasing tendency, Flanagan's plots reach a constant 100% for large formant frequency shifts due to the fact that subjects in his tests were asked to merely say if they perceived the two speech sounds corresponding to unperturbed and perturbed sets of formants as being different rather than to quantify the amount of quality difference they perceived between the two sounds.

The effectiveness of the weighting $A(\omega)$ is particularly apparent in the low frequency region. Fig. 6 shows plots of spectral distance with and without this weighting, other conditions being the same, for the case when the first formant is shifted about 300 Hz. The unweighted measure gives a slight asymmetry but in the wrong sense, Fig. 6(a), while the weighted measure produces the right asymmetry as shown by Fig. 6(b).

A disadvantage of the distance measures presented in this section is that they are dependent on the energy of the spectra. (Notice that distance measures which are functions of only the ratio of spectra do not suffer from this disadvantage.) With energy dependent measures, comparison of spectral distances obtained, for instance, for different analysis situations can be meaningfully done only after suitably scaling the distance values. A reasonable condition

to impose on such scaling is that the spectral distances corresponding to the formant frequency difference limens at the different frequencies be approximately equal. This will be our next step in refining the class of perceptually consistent spectral distance measures that we suggested above.

6. Conclusions

We have reported preliminary results of an ongoing work on perceptually consistent spectral distance measures. Our experience has been that GM normalization works better than AM normalization inasmuch as one is looking for sensitivity to interaction of formants. The results we have presented in this paper show that the distance is best defined in terms of the difference in the (linear) spectral values. Besides continuing our investigation reported here, we plan to use the developed measures in several applications.

Acknowledgment

This work was supported by the Information Processing Techniques Branch of the Advanced Research Projects Agency.

References

1. A. Ichikawa, Y. Nakano and K. Nakata, "Evaluation of Various Parameter Sets in Spoken Digits Recognition," IEEE Trans. AU, 202-209, June 1973.
2. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. ASSP, 67-72, Feb. 1975.
3. J. Makhoul."Linear Prediction in Automatic Speech Recognition," in Speech Recognition, R. Reddy (Ed.), New York: Academic Press, 1975.
4. B.S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," JASA, 1304-1312, June 1974.
5. S.F. Boll, "Waveform Comparison Using the Linear Prediction Residual," Comp. Science Dept., Univ. Utah, 1975.
6. D. T. Magill,"Adaptive Speech Compression for Packet Communication Systems," Proc. Nat'l Telecommun. Conf., Nov. 1973.
7. J. Makhoul, R. Viswanathan, L. Cosell and W. Russell, Natural Communication with Computers, Vol. II, Speech Compression Research at BBN, Rept.No. 2976, Dec. 1974.
8. J. Makhoul, R. Viswanathan and W. Russell, "A Framework for the Objective Evaluation of Vocoder Speech Quality," Internat'l Conf. ASSP, Philadelphia, April 1976.
9. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. ASSP, 309-321, June 1975.
10. J.L. Flanagan, "A Difference Limen for Vowel Formant Frequency," JASA, 613-617, May 1955.
11. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, 561-580, April 1975.
12. A.H. Gray, Jr. and J.D. Markel, "Distance Measures for Speech Processing," Submitted for publication in IEEE Trans. ASSP.
13. BBN Quarterly Progress Report, Command and Control Related Computer Technology, BBN Rept. No. 3122, Sept. 1975.
14. S.S. Stevens, "Perceived Level of Noise by Mark VII and Decibels (E)," JASA, 575-600, Feb. 1972.

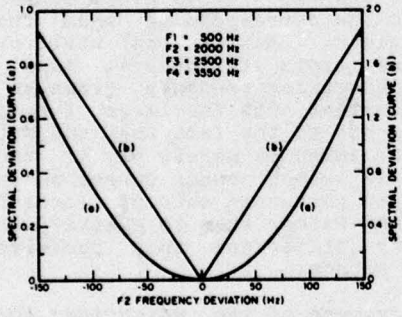


Fig. 1

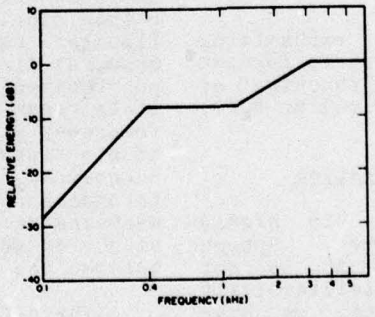


Fig. 2

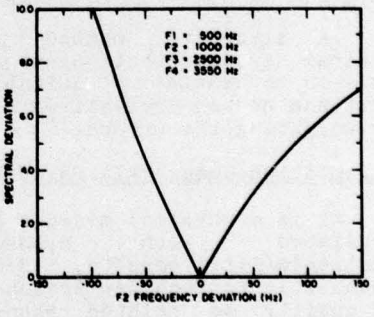


Fig. 3

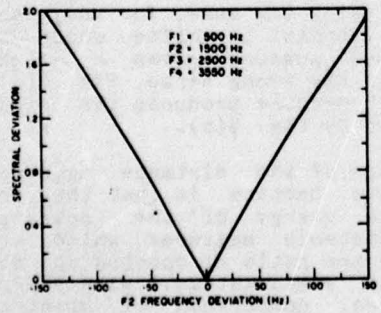


Fig. 4

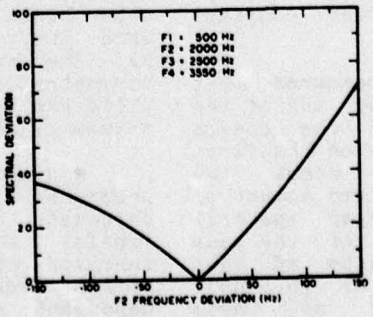


Fig. 5

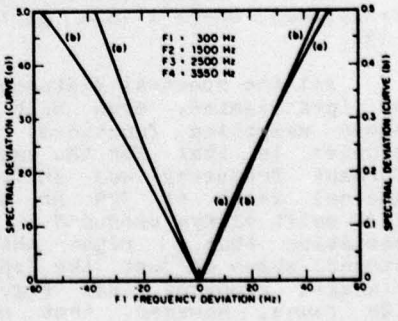


Fig. 6

APPENDIX 14

OBJECTIVE SPEECH QUALITY EVALUATION OF NARROWBAND LPC VOCODERS

(Paper to be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK, April 1978.)

OBJECTIVE SPEECH QUALITY EVALUATION OF NARROWBAND LPC VOCODERS

R. Viswanathan, W. Russell and J. Makhoul

Bolt Beranek and Newman Inc.
Cambridge, Mass. 02138

ABSTRACT

Several methods are presented for the objective speech quality evaluation of narrowband LPC vocoders, based on a framework that we proposed at the 1976 ICASSP conference. In each method, the error in short-term spectral behavior between vocoded speech and the original is computed once every 10 ms. These errors are appropriately weighted and averaged over an utterance to produce a single objective score. Several short-term error measures, and time-weighting and averaging techniques are investigated. We evaluate the objective methods by correlating the resulting objective scores with formal subjective speech quality judgments. High correlations obtained indicate the usefulness of these methods.

1. INTRODUCTION

Quality assessment of vocoded speech is often performed to determine the user acceptance of a vocoder, or to compare the performance of competing vocoder types, or to evaluate the different choices of a given vocoder's design parameters. Procedures used for speech quality measurement are either subjective or objective, depending upon whether or not they make use of subjective judgments from human listeners. Subjective procedures require extensive testing with human listeners, which is expensive in terms of both time and money. On the other hand, objective measures would enable evaluation to be done by computer as well as ensure uniformity in speech quality evaluation. Also, objective measures can be incorporated into the design of better quality vocoders. Of course, the validity of any objective procedure must first be established by comparing its results against subjective judgments.

While there exist a few subjective procedures, relatively little work has been done to develop objective procedures. The purpose of this paper is to report on several objective measures for speech quality assessment of narrowband linear predictive (LPC) vocoders. We have developed these measures based on a general framework for objective speech quality evaluation that we presented at the 1976 ICASSP conference. A specific objective procedure presented also at that conference [3] falls within this general framework.

2. A FRAMEWORK FOR OBJECTIVE SPEECH QUALITY EVALUATION

Any objective measure for the evaluation of vocoded speech quality must be a function of the vocoder transmission parameters and must somehow relate to perception. Using this and a number of other detailed arguments, we formulated in [1] a general framework for objective speech quality evaluation, based on the following reasonable assumptions:

1) Speech synthesized from unquantized vocoder parameters extracted every 10 ms, is of very good quality compared to the original speech.
2) Except for pitch and gain, the fidelity of the short-time speech spectrum is the principal determiner of quality.
3) The spectrum is uniquely defined by the linear prediction filter parameters.

The first assumption gives us an anchor point, defined in terms of the unquantized vocoder parameters, against which to compare quantized realizations of the same utterance. The second and third assumptions relate the filter parameters to speech quality. In the above formulation, we have implicitly made an important assumption that the vocoder under evaluation extracts pitch values and voiced/unvoiced decisions without any error. Although the formulation may be extended to cover voice-excited and residual-excited LPC vocoders as well as vocoders other than LPC (e.g., channel vocoders) [1], here we consider its application exclusively to pitch-excited (narrowband) LPC vocoders.

In the above framework the problem of objective quality evaluation is reduced to the following two steps: 1) For each 10 ms frame, compute an objective error as the distance or deviation between the spectrum corresponding to the unquantized LPC parameters and the spectrum corresponding to the quantized and interpolated LPC parameters (interpolation is required if the vocoder's transmission frame rate is lower than 100 frames/sec.); and 2) combine all the frame errors thus computed within a speech utterance into one number, which becomes the objective speech quality score. Methods for carrying out the two tasks are presented in Sections 3-5.

3. FRAME SPECTRAL ERROR MEASURES

The power spectrum of linear prediction all-pole filter is given by

$$P(\omega) = G^2/S(\omega) = R_0 V_p / \left|1 + \sum_{k=1}^{p} a_k e^{-j\omega k}\right|^2 \qquad (1)$$

where G is the filter gain, R_0 is the speech signal energy, V_p is the normalized preediction error, $S(\omega)$ is the spectrum of the inverse filter, a_k are the predictor coefficients, and p is the order of the linear predictor. To compute objective frame error, we require a measure of distance between the reference spectrum $P_1(\omega)$ (unquantized parameters) and the vocoded speech spectrum $P_2(\omega)$ (quantized and interpolated parameters). Although there are a number of distance measures available [2], we consider in this paper three distance measures denoted below as d1, d2 and d3.

CH1285-6/78/0000-0591$00.75@1978IEEE

For d1 and d2, the distance between the spectra P_1 and P_2 is computed in three steps as follows:

a) Normalize the two spectra by making them have the same (unity) geometric mean (i.e., the areas under the log spectra are made equal);

b) Determine the error at each frequency either as the magnitude of the difference in linear spectral amplitudes of the two normalized spectra (d1) or as the magnitude of the difference in their log spectral amplitudes (d2); and

c) Compute a suitable norm of this error function.

Notice that the geometric mean of the power spectrum in (1) is $V_p R_0$. The two measures d1 and d2 are given below [2,4]:

$$d1 = \frac{\int_0^{\pi} A(\omega)\,|S_1^{-1}(\omega) - S_2^{-1}(\omega)|\,d\omega}{\int_0^{\pi} A(\omega)\,S_1^{-1}(\omega)\,d\omega} \qquad (2)$$

$$d2 = \left[\frac{1}{\pi}\int_0^{\pi} |\log S_1(\omega) - \log S_2(\omega)|^2 d\omega\right]^{\frac{1}{2}} \qquad (3)$$

where $A(\omega)$ is a smoothed version [4] of the weighting function originally introduced by S.S. Stevens for measuring the perceived loudness [2]. The distance measure d1 is perceptually consistent in the sense that it produces results consistent with published subjective perceptual results on formant frequency difference limens [2,4], while d2 is not perceptually consistent.

The third distance measure d3 measures the absolute deviation in the log area ratios (LARs) g_k, which are uniquely related to the predictor coefficients a_k, and which possess flat or uniform spectral sensitivity as well as other desirable properties [5]:

$$d3 = \frac{1}{p}\sum_{k=1}^{p} |g_{1k} - g_{2k}| \quad , \qquad (4)$$

where the two sets of LARs correspond to the two linear predictors. Since we deal with LPC vocoders that employ LARs as transmission parameters, they are readily available for evaluating d3 "inside" the vocoder. Of the above three measures, clearly d3 is least expensive to compute.

4. TIME-WEIGHTING OF FRAME ERRORS

The task of combining the frame errors E(i) into a single speech quality score involves first weighting the frame errors with a suitable time-weighting function W(i) to reflect the relative importance of the individual frames to perceived speech quality, and then averaging the weighted frame errors E(i)W(i). Below, we describe two time-weighting methods that we have investigated.

Energy Weighting

In this method, we make the reasonable assumption that frame errors in low energy regions of an utterance have a smaller influence on quality judgments than those in high energy regions. For example, large changes in the spectrum may not be detected by the listener if the total energy in the spectrum is low. We considered the weighting as a function of the frame speech signal energy per sample, computed over an interval of 20 ms and expressed in decibels. We have investigated linear (W1) and piecewise-linear (W2) weighting functions. These are depicted in Fig. 1. The piecewise-linear function shown is less drastic than the linear function in that it deemphasizes frame errors in the low energy region, but has only a slight influence on all other frame errors.

Dynamic Fidelity Weighting

Another consideration relevant to speech quality is the rate at which speech characteristics change in time. This rate varies in time in accordance with the sequence of speech sounds being uttered. Since it is reasonable to assume that a rate of LPC parameter extraction of 100 frames/sec (fps) should be adequate to track all perceptually important speech events, we chose the anchor system as having 100 fps LPC data (Section 2). However, in the case of slowly varying speech materials (e.g., JB1, see Section 6), the actual rate of change of speech characteristics is substantially lower than that in normal speech. This means that parameter extraction at rates much less than 100 fps can adequately represent the slowly varying speech. This poses two problems with respect to the choice of our anchor system. First, the objective speech quality measure computed based on the above anchor would generally yield lower error when the transmission frame rate of the vocoder under evaluation is closer to 100 fps. This is because when the vocoder's frame rate is closer to the anchor system's frame rate, frame error computation involves fewer parameter errors due to interpolation, which are being substituted by quantization errors, and these are generally smaller than interpolation errors. Therefore, for slowly varying speech the objective measure would overestimate the vocoded speech quality. Secondly, subjective speech quality tests have clearly shown that a 100 fps LPC vocoder produces inferior speech quality (characterized as having a "wobble" quality) when processing slowly varying speech, compared to a 50 fps vocoder [6]. To overcome these problems, we redefine the anchor system based on a functional perceptual model (PM) of speech that two of the authors have recently formulated [6]. In this model, it is postulated that (1) Speech can be represented in terms of LPC (or other) parameters extracted at a minimal set of perceptually significant frames, not necessarily equally spaced, and (2) Between any two such frames, the parameters vary linearly. An automatic scheme has been developed to determine or "mark" the location of the perceptually significant frames [6]. The PM-based anchor system is therefore characterized by the LPC parameters (actually, LARs) of the frames marked by this scheme and the linearly interpolated parameter values for the unmarked frames. We have presented this modification in this section, since it may be viewed as an implicit time-weighting method. In addition, we have investigated an explicit time-weighting in which frame errors for the marked frames are weighted with unity, while other frame errors are weighted with a fraction depending on the duration of the transmission interval to which they belong. In the interest of keeping the presentation of results in Section 6 brief, we assume unity weighting for all marked and unmarked frames.

5. TIME-AVERAGE OF WEIGHTED FRAME ERRORS

The final step in formulating an objective speech quality measure is to specify how the weighted frame errors W(i)E(i) are combined into one number. One obvious method is to use the weighted r-th mean of all the (say, L) frame errors over the whole utterance:

$$\left[\sum_{i=1}^{L} W(i) \, [E(i)]^r \Big/ \sum_{i=1}^{L} W(i) \right]^{1/r} \qquad (5)$$

The simplest average of this type is the arithmetic mean with r=1; this average is denoted by E1. Two other averages E3 and E4 that we have used are described below.

Subjective quality judgments of an utterance are greatly influenced by the presence of even one or two places of large errors such as those that are perceived as pops or glitches. An overall average such as E1 may not portray such influences, especially if most of the remaining frame errors are small. The above r-th mean with a large r would emphasize large frame errors. An alternate method of achieving this selectivity to large errors is to consider the arithmetic mean over only a specified number of the largest frame errors. We define a measure E2 which is the average over the top 10% of the frame errors. A two-term composite average measure E3 is obtained as the sum of E1 and E2. A different composite average is motivated by the consideration that the number of large frame errors which influence perceived quality of an utterance may vary from one vocoded version to another. This suggests that E2 may be replaced by a selective average that is carried out over a variable percentage of top frame errors, or alternately E2 may be multiplied with a variable weight and then added to E1. We denote this weighted composite average by E4:

$$E4 = E1 + \gamma E2 \qquad (6)$$

In our experimental investigations, we obtained significant improvements in correlation scores when we chose the following exponential weight γ:

$$\gamma = k1 \, e^{k2 \cdot \alpha_3} \qquad (7)$$

where k1 and k2 are constants, and α_3 is the "skewness" of the frame error distribution over the whole utterance, defined by

$$\alpha_3 = \frac{1}{L} \sum_{i=1}^{L} [E(i) - E1]^3 / \sigma_E^3 \ , \qquad (8)$$

with σ_E being the standard deviation. Use of k2=-1 was found to improve the performance of the objective measures as determined by correlation against subjective judgments.

6. CORRELATION AGAINST SUBJECTIVE JUDGMENTS

With three frame error measures d1-d3 (Section 3), two energy weighting functions W1 and W2 (Section 4), the perceptual-model-based dynamic fidelity weighting scheme (Section 4), and three time-average measures E1, E3, and E4 (Section 5), and considering different values for the constants that figure in some of the above items, we get a large variety of possible objective speech quality measures. To evaluate the effectiveness of a given objective quality measure, we correlate the objective scores that the measure produces for an utterance processed through a range of LPC vocoder systems against the corresponding subjective quality judgments. Notice that the objective scores for the various vocoded versions of an utterance are all based on the same anchor, and hence the scores can be directly compared to infer quality differences between different vocoders in processing that utterance.

We compute two types of correlation between the objective and subjective data: (1) regular correlation (or simply correlation); and (2) rank order correlation. For the second type, two sets of ranks are first assigned to vocoders under study using separately objective and subjective data, and then regular correlation is computed between the two sets of ranks. Therefore, rank order correlation is useful in examining how well an objective measure would order vocoders in terms of perceived (subjective) speech quality. On the other hand, unlike the rank order correlation, regular correlation is sensitive to the actual extents of vocoder quality differences.

Below, we first briefly discuss the subjective speech quality rating that was collected in a recent study [7], and then present highlights of the results obtained by correlating objective quality data against this subjective data.

Subjective Data Base

Stimuli for the subjective rating study [7] were made by passing 7 sentences through each of 49 fixed-rate LPC vocoders. The transmission bit rates for those vocoders ranged from 1267 bps to 8700 bps. The different vocoder systems were obtained by varying, in a factorial manner, the LPC order (13, 11, 9 and 8 poles), the LAR quantization step size (0.5, 1 and 2 dB) and the transmission frame rate (100, 67, 50 and 33 fps). The 49th vocoder had 13 poles, 0.25 dB step size, and 100 fps frame rate. A 110 kbps PCM speech (11-bit samples at 10 kHz), which was the input to the vocoders, was also included to act as an undegraded anchor. Nine subjects rated speech quality degradation on a scale of 0-100. The rating scores averaged over the nine subjects gave the subjective data for our correlation study. To keep the overall task manageable (in view of the large variety of objective measures we were considering), we chose a subset of 22 vocoders (all the 12 13-pole systems, and 5 each of 11-pole and 8-pole systems) and 5 sentences given in Table 1.

Correlation Results

We ran extensive correlation experiments for several purposes: 1) to use the correlation scores as a means of choosing the parameters of the time-weighting and time-averaging schemes discussed above; 2) to study the effect of incorporating into an objective evaluation measure any one or group of the different time-weighting and time-averaging schemes; and, 3) to pick that subset of these schemes which, for a given frame error measure, maximizes the correlation scores on the average. In short, correlation against subjective data served as the primary vehicle for judging the effectiveness of an objective quality measure. Results obtained from these correlation studies are numerous and complex due to the interactions between the different time-weighting and averaging elements as well as the profound effect of speech material and speakers. Below, we provide important highlights of these results. Since the vocoder input speech used in this study had a 5 kHz bandwidth, we employed a 14-th order anchor system in computing the objective quality scores.

a) Correlation scores obtained for male speech (JB1, JB5 and DK6) were generally higher than those for female speech (AR4 and RS6). (See further below.)

b) The energy weighting function W1 or W2 and the PM-based implicit time-weighting method produced in general higher correlations, although the two methods did not always reinforce each other.

c) By and large the averaging method E4 is superior to E1 and E3.

d) For the frame error measures d1-d3, we found the appropriate time-weighting and averaging methods so as to secure on the average maximum correlation scores for the 5 utterances we considered. The resulting objective quality measures M1-M3 are described in Table 2, and their correlation scores are given in Table 3. For both M2 and M3, PM-based weighting and no (or unity) energy weighting were chosen, while for M1 linear energy weighting W1 and no PM-based weighting were preferred. This may partly be due to the fact that the automatic PM scheme already uses (linear) energy weighting [6].

e) The correlation scores given in Table 3 range from 0.685 to 0.947, and are all highly significant. Note that for 22 "measurements" corresponding to 22 vocoders a significance level of better than 0.001 is achieved with a correlation score of only 0.652.

f) The measure M2 based on the rms log spectral error and the measure M3 based on the LAR error were found to behave quite similarly. Since all three quality measures produced good correlation results, we recommend the use of M3 as it is the least expensive of the three computationally.

Attempts to Improve Objective Quality Measurement of Female Speech

As mentioned above, and as shown in Table 3, all the reported objective measures yielded generally lower correlation scores for female speech than for male speech. Also, in choosing the components of the objective measures M1-M3 given in Table 2 we did not make use of the correlation scores computed for AR4 and RS6, since the scores varied relatively spuriously and did not indicate any clear choice. One reason for these problems is that the 22 vocoders considered in our correlation study drew in general more clustered subjective ratings for AR4 and RS6 than for JB1, JB5 and DK6. A second reason (somewhat related to the first) is that subjective rating scores for the utterances from female speakers were relatively constant over the range of the LPC order considered (8-13 poles); in contrast, the rating scores for male speakers exhibited a wide range of variation. Also, the subjective rating data for female speech had several examples where a vocoder with a lower order drew a better rating than another with a higher order, with the quantization step size and frame rate being the same for the two vocoders. These considerations suggest that the order of the anchor system may be varied as a function of the average fundamental of the speaker over the whole utterance. By choosing the anchor system's order as 12 poles for AR4 and 10 poles for RS6, we obtained definite increases in the correlation scores, although the scores still remained substantially lower (especially for RS6) than those obtained for male speech.

ACKNOWLEDGMENT

The research was sponsored by the Information Processing Techniques branch of the Advanced Research Projects Agency.

REFERENCES

1. J. Makhoul, R. Viswanathan and W. Russell, "A Framework for the Objective Evaluation of Vocoder Speech Quality," Proc. 1976 ICASSP, pp. 103-106, April 1976.
2. R. Viswanathan, J. Makhoul and W. Russell, "Towards Perceptually Consistent Measures of Spectral Distance," Proc. 1976 ICASSP, pp. 485-488, April 1976.
3. S. Meister and R.H. Wiggins, "Quality Comparison Measure for Linear Predictive Systems," Proc. 1976 ICASSP, pp. 107-109, April 1976.
4. BBN Quarterly Progress Report on Command and Control Related Computer Technology, BBN Report No. 3325, June 1976.
5. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. ASSP, pp. 309-321, June 1975.
6. R. Viswanathan, J. Makhoul and R. Wicke, "The Application of a Functional Perceptual Model of Speech to Variable-Rate LPC Systems," Proc. 1977 ICASSP, pp. 219-222, May 1977.
7. A.W.F. Huggins, R. Viswanathan and J. Makhoul, "Quality Ratings of LPC Vocoders: Effects of Number of Poles, Quantization and Frame Rate," Proc. 1977 ICASSP, pp. 413-416, May 1977.

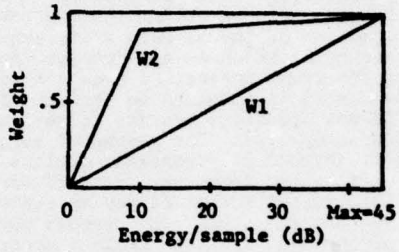


Fig. 1 Energy weighting functions.

ID	F0	Sentence
JB1	119	Why were you away a year, Roy?
AR4	165	Which tea-party did Baker go to?
JB5	124	The little blankets lay around on the floor.
DK6	97	The trouble with swimming
RS6	193	is that you can drown.

Table 1. The five stimulus sentences, with the speaker's average fundamental frequency in Hz.

Quality Measure	Frame Error	Energy Weighting	PM-based Weighting	Time Average
M1	d1	Linear, W1	No	E4
M2	d2	None	Yes	E3
M3	d3	None	Yes	E3

Table 2. Description of 3 objective speech quality measures.

Sentence		Objective Speech Quality Measure		
		M1	M2	M3
Males	JB1	.916(.699)	.927(.808)	.905(.872)
	JB5	.929(.928)	.868(.823)	.920(.899)
	DK6	.947(.848)	.887(.825)	.909(.684)
Females	AR4	.853(.807)	.885(.866)	.928(.879)
	RS6	.716(.718)	.812(.817)	.685(.653)

Table 3. Correlation scores obtained for 3 objective measures. Values in parentheses are rank order correlation scores.